This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

THIS PAGE BLANK (USPTO)

(12) UK Patent Application (19) GB (11) 2 336 697 (13) A

(43) Date of A Publication 27.10.1999

(21) Application No 9808801.6

(22) Date of Filing 24.04.1998

(71) Applicant(s)
The Dialog Corporation Plc
(Incorporated in the United Kingdom)
The Communications Building, 48 Leicester Square,
LONDON, WC2H 7DB, United Kingdom

(72) Inventor(s)
Rachel Hammond
Llewelyn Ignazio Fernandes

(74) Agent and/or Address for Service
Atkinson & Co
First Floor, Unit A, The Technology Park,
60 Shirland Lane, SHEFFIELD, S9 3PA,
United Kingdom

(51) INT CL^6
G06F 17/30

(52) UK CL (Edition Q)
G4A AUDB

(56) Documents Cited
US 5537586 A US 5428778 A

(58) Field of Search
UK CL (Edition P) G4A AUDB
INT CL^6 G06F 17/30
Online: WPI,USA

(54) Abstract Title
Alerting user-processing sites as to the availability of information

(57) Data files are received by a central processing system (104) and these files analysed to determine whether they contain information which is relevant to user-specified characteristics. On detecting such a condition, an alert signal is supplied to the respective user (115). The incoming data files are analysed with r spect to common data characteristics to generate common category associations. The data files are then processed with respect to user-specific data characteristics. The user-specific data characteristics include examples of the common data characteristics and the specific processing procedures make use of the previously defined common category associations.

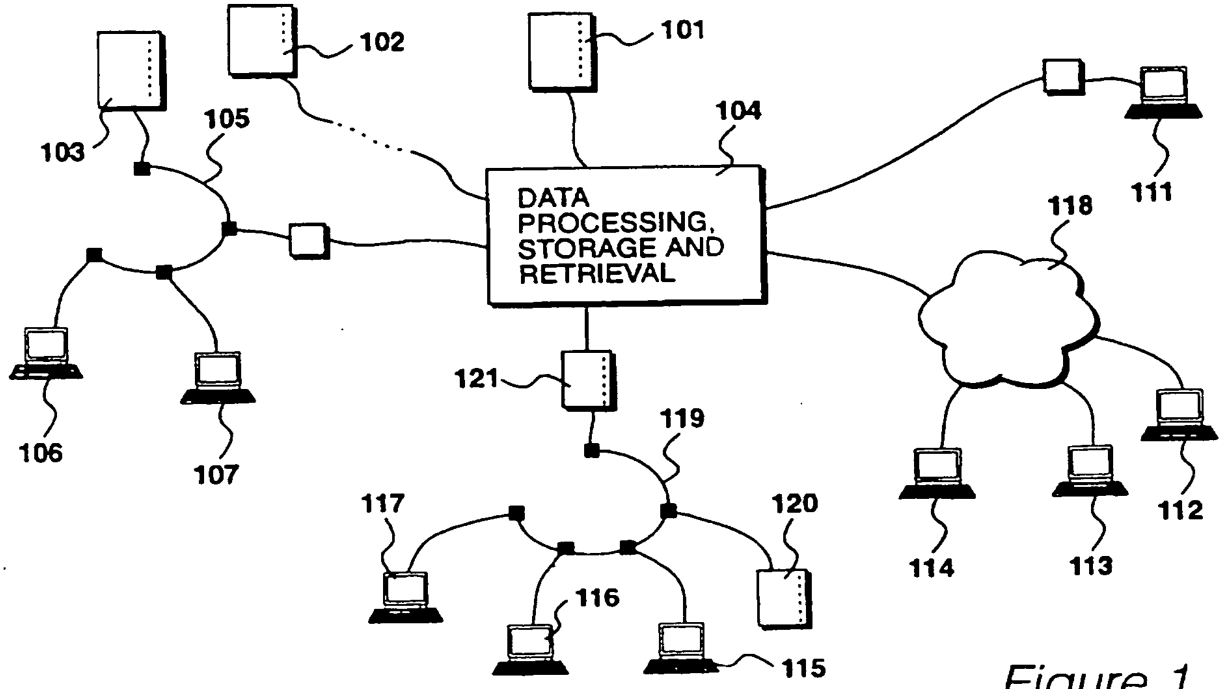


Figure 1

GB 2 336 697 A

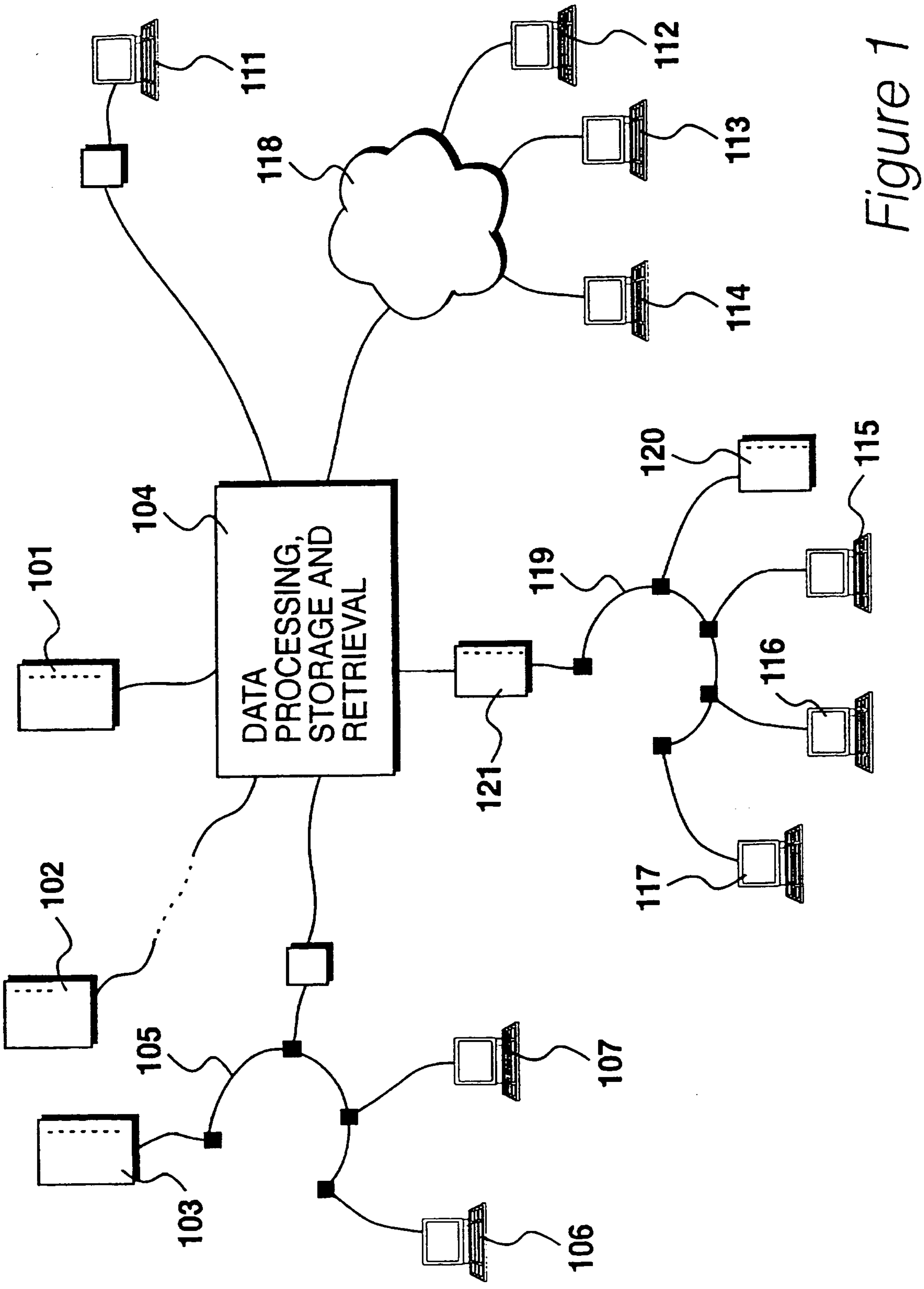

111
112
118
113
114
104
DATA PROCESSING, STORAGE AND RETRIEVAL
101
102
121
119
120
115
116
117
105
103
107
106

Figure 1

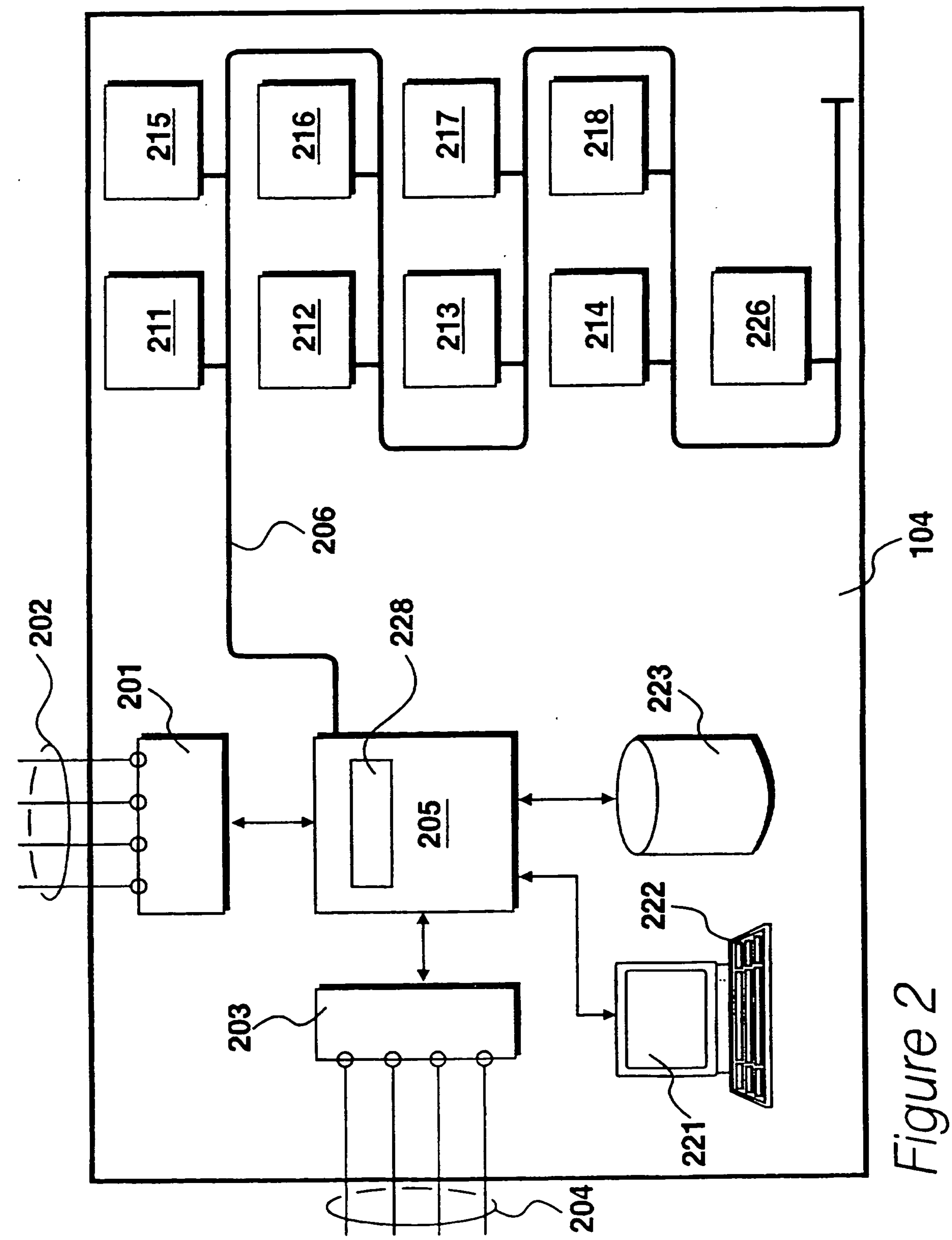

202
201
228
205
206
223
222
221
203
204
104
211
212
213
214
226
215
216
217
218

Figure 2

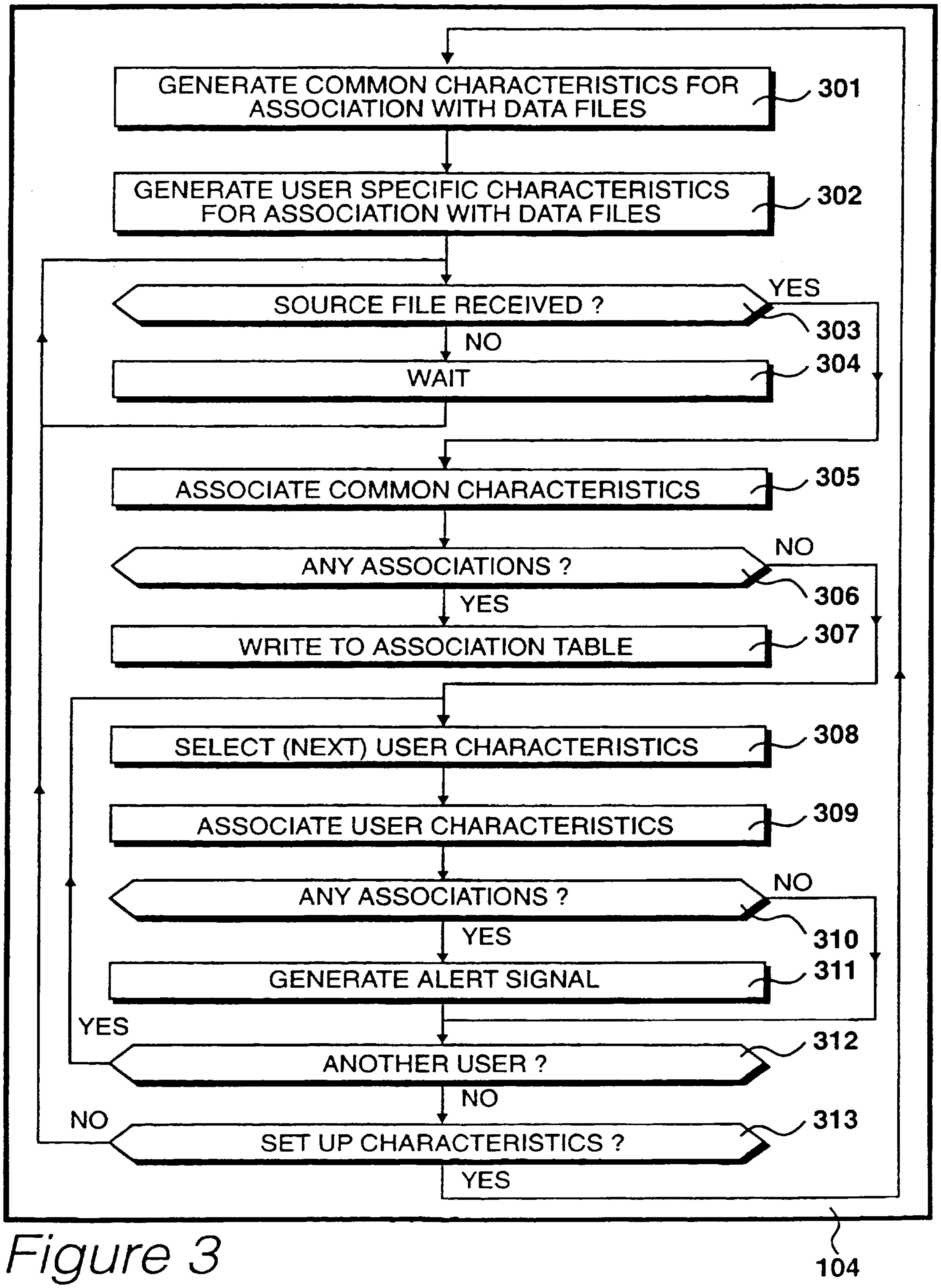

GENERATE COMMON CHARACTERISTICS FOR ASSOCIATION WITH DATA FILES
301
GENERATE USER SPECIFIC CHARACTERISTICS FOR ASSOCIATION WITH DATA FILES
302
SOURCE FILE RECEIVED ?
YES
303
NO
WAIT
304
ASSOCIATE COMMON CHARACTERISTICS
305
ANY ASSOCIATIONS ?
NO
306
YES
WRITE TO ASSOCIATION TABLE
307
SELECT (NEXT) USER CHARACTERISTICS
308
ASSOCIATE USER CHARACTERISTICS
309
ANY ASSOCIATIONS ?
NO
310
YES
GENERATE ALERT SIGNAL
311
YES
ANOTHER USER ?
312
NO
NO
SET UP CHARACTERISTICS ?
313
YES
104

Figure 3

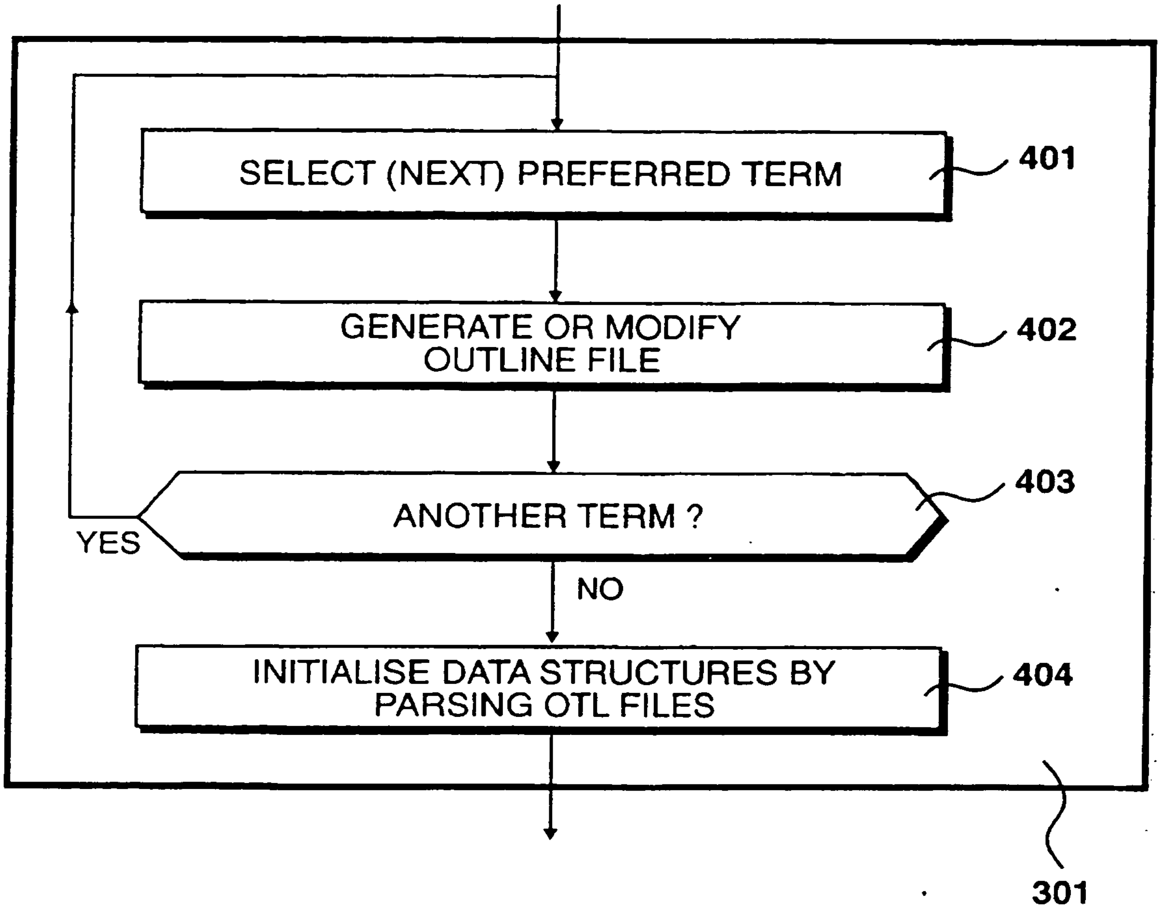

SELECT (NEXT) PREFERRED TERM
401
GENERATE OR MODIFY OUTLINE FILE
402
ANOTHER TERM ?
403
YES
NO
INITIALISE DATA STRUCTURES BY PARSING OTL FILES
404
301

Figure 4

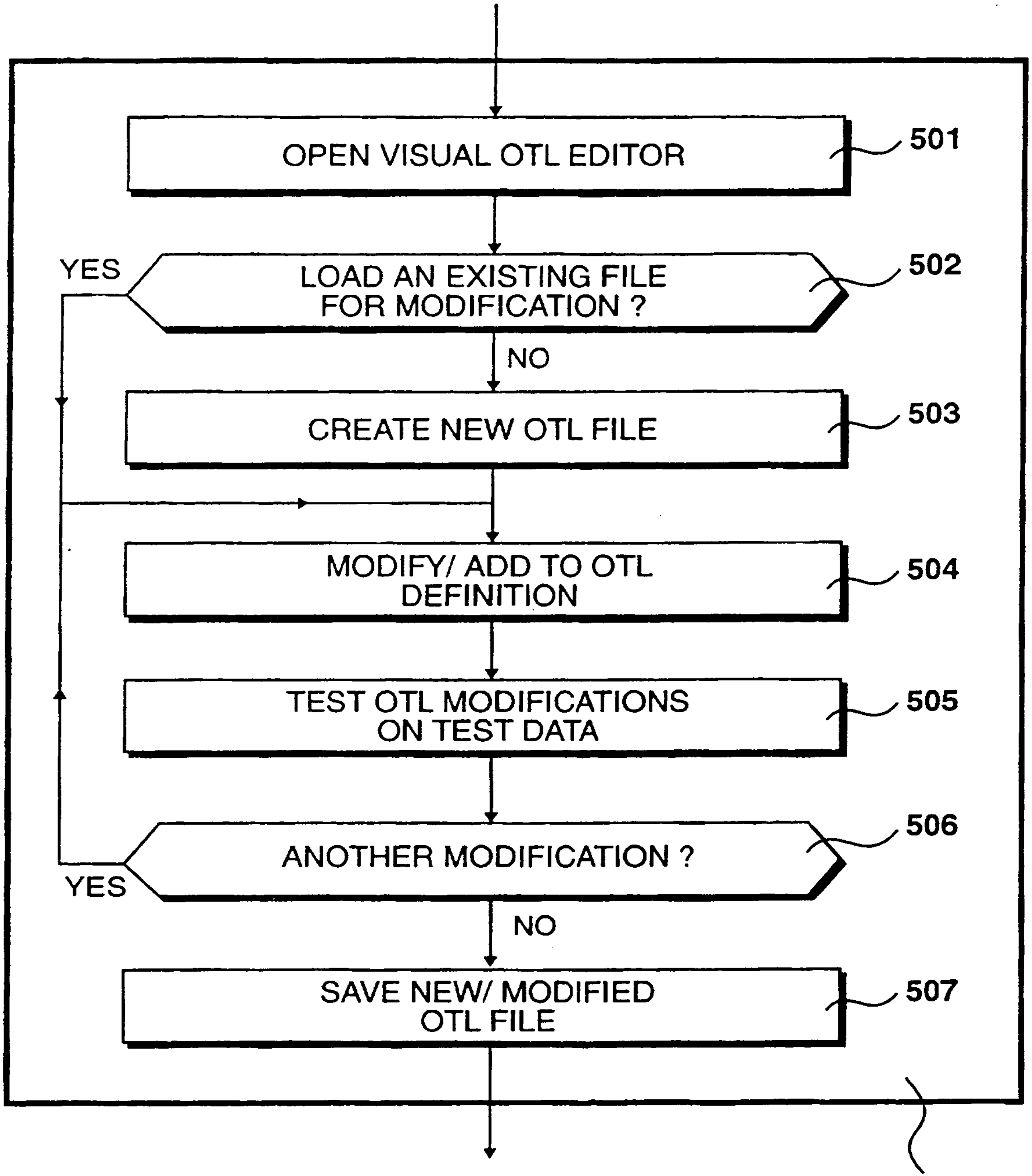
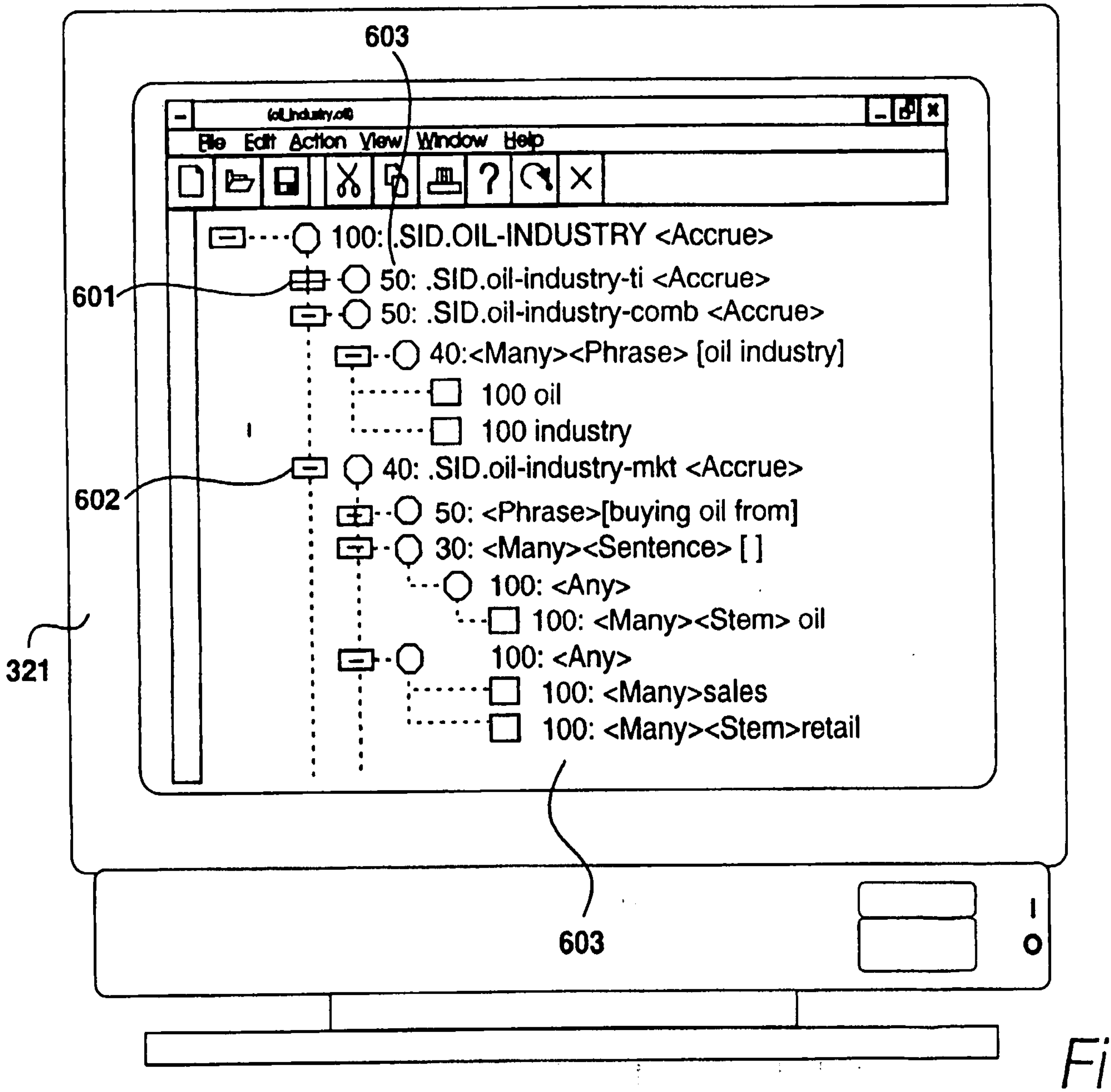
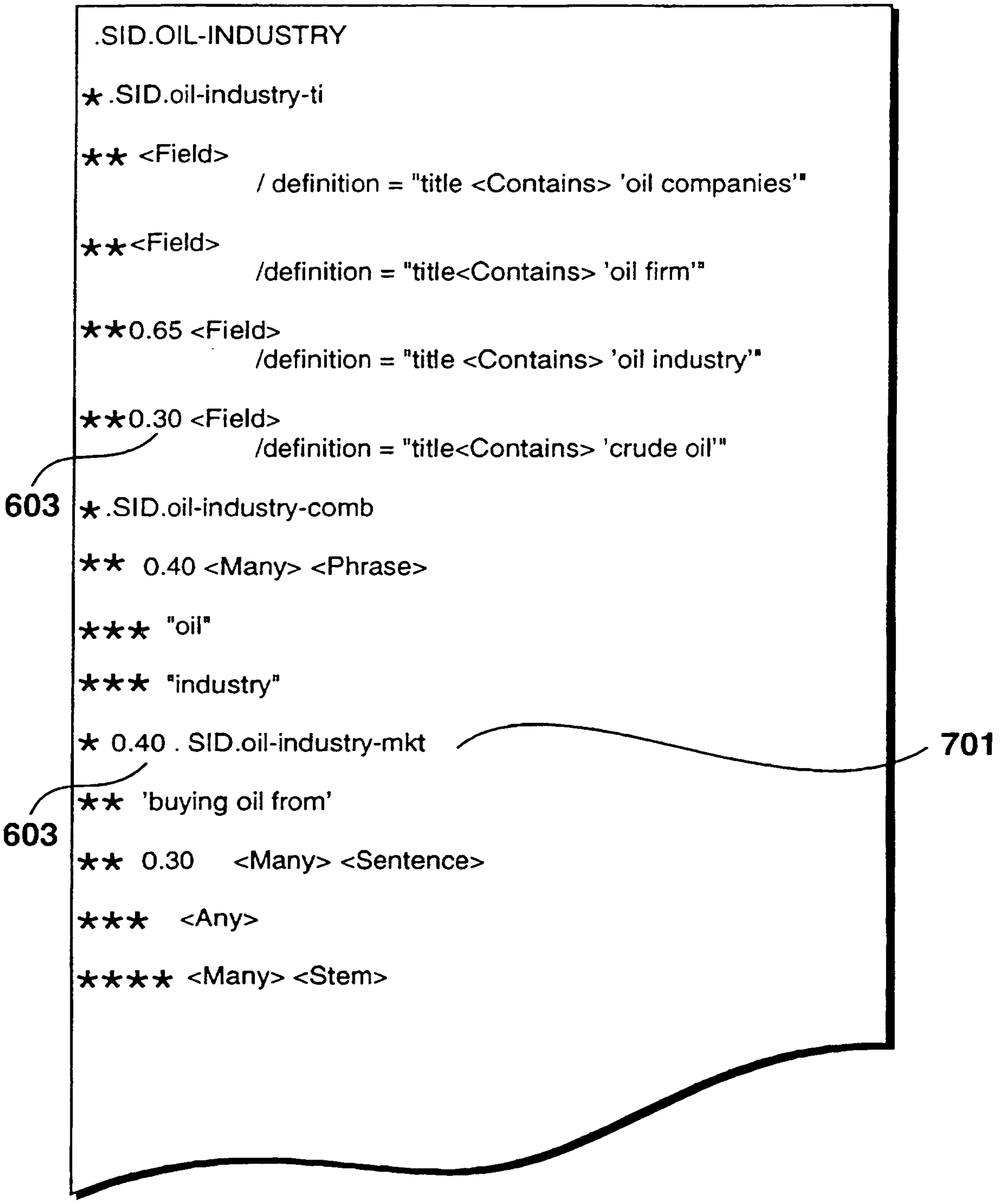

OPEN VISUAL OTL EDITOR
501
LOAD AN EXISTING FILE FOR MODIFICATION ?
502
YES
NO
CREATE NEW OTL FILE
503
MODIFY/ ADD TO OTL DEFINITION
504
TEST OTL MODIFICATIONS ON TEST DATA
505
ANOTHER MODIFICATION ?
506
YES
NO
SAVE NEW/ MODIFIED OTL FILE
507
402

Figure 5



```
100: .SID.OIL-INDUSTRY <Accrue>
    50: .SID.oil-industry-ti <Accrue>
    50: .SID.oil-industry-comb <Accrue>
        40:<Many><Phrase> [oil industry]
            100 oil
            100 industry
    40: .SID.oil-industry-mkt <Accrue>
        50: <Phrase>[buying oil from]
        30: <Many><Sentence> [ ]
            100: <Any>
                100: <Many><Stem> oil
        100: <Any>
            100: <Many>sales
            100: <Many><Stem>retail
```

Figure 6



```
.SID.OIL-INDUSTRY

*.SID.oil-industry-ti

** <Field>
            / definition = "title <Contains> 'oil companies'"

**<Field>
            /definition = "title<Contains> 'oil firm'"

**0.65 <Field>
            /definition = "title <Contains> 'oil industry'"

**0.30 <Field>
            /definition = "title<Contains> 'crude oil'"

*.SID.oil-industry-comb

**  0.40 <Many> <Phrase>

***  "oil"

***  "industry"

* 0.40 . SID.oil-industry-mkt

**  'buying oil from'

**  0.30     <Many> <Sentence>

***    <Any>

**** <Many> <Stem>
```

Figure 7

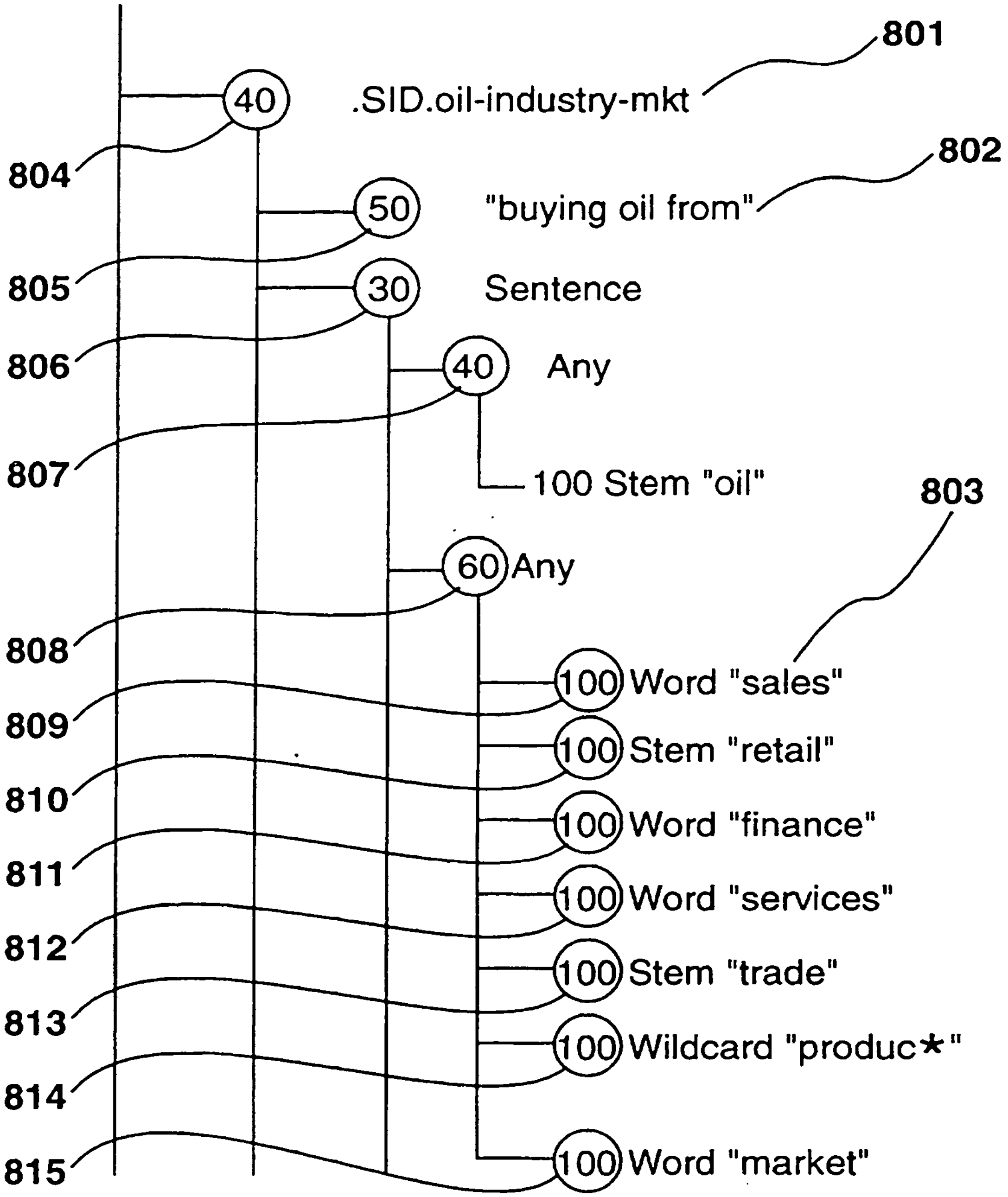

801
40
.SID.oil-industry-mkt
804
802
50
"buying oil from"
805
30
Sentence
806
40
Any
807
100 Stem "oil"
803
60
Any
808
100
Word "sales"
809
100
Stem "retail"
810
100
Word "finance"
811
100
Word "services"
812
100
Stem "trade"
813
100
Wildcard "produc*"
814
100
Word "market"
815

Figure 8

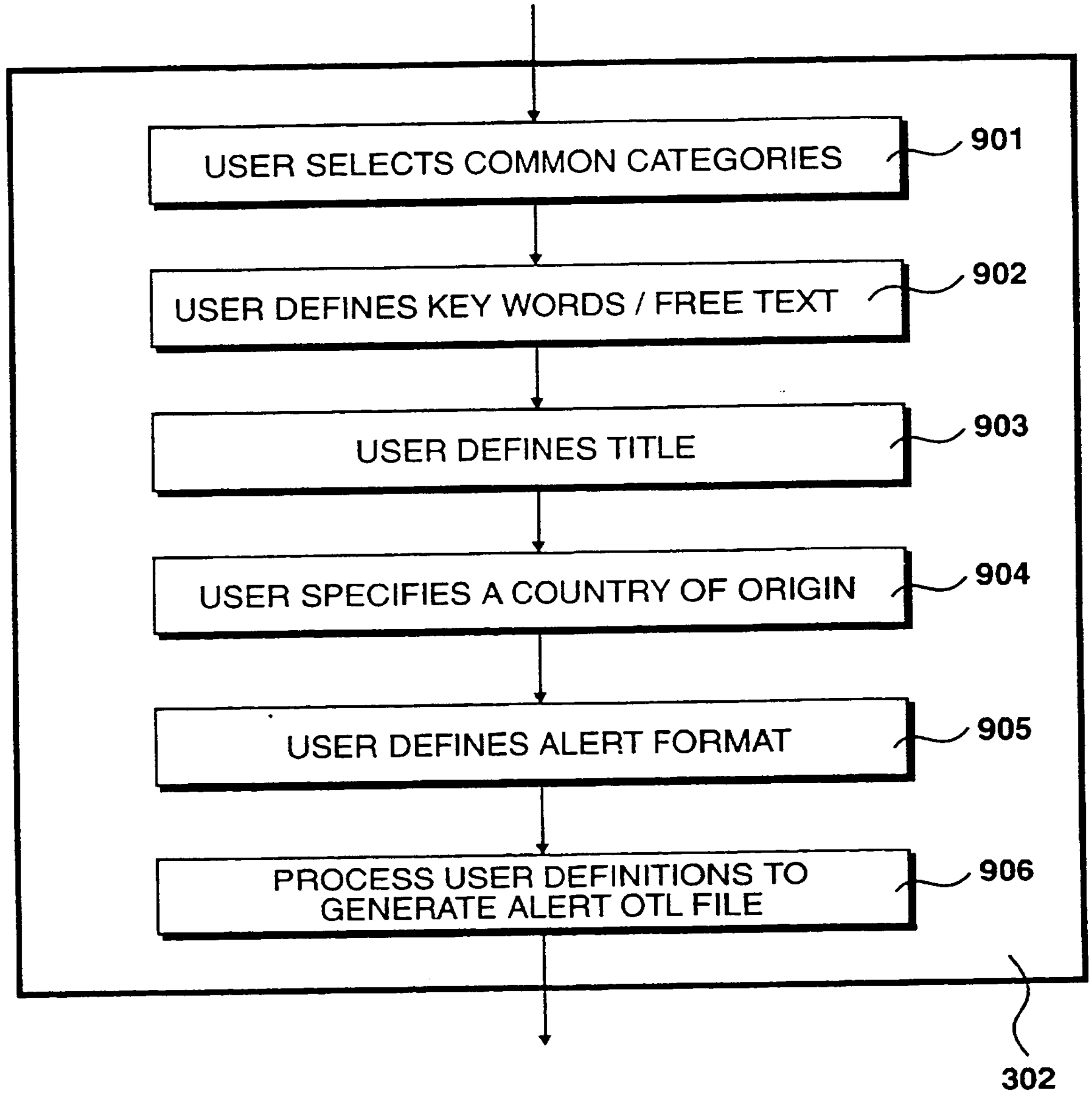

USER SELECTS COMMON CATEGORIES
901
USER DEFINES KEY WORDS / FREE TEXT
902
USER DEFINES TITLE
903
USER SPECIFIES A COUNTRY OF ORIGIN
904
USER DEFINES ALERT FORMAT
905
PROCESS USER DEFINITIONS TO GENERATE ALERT OTL FILE
906
302

Figure 9

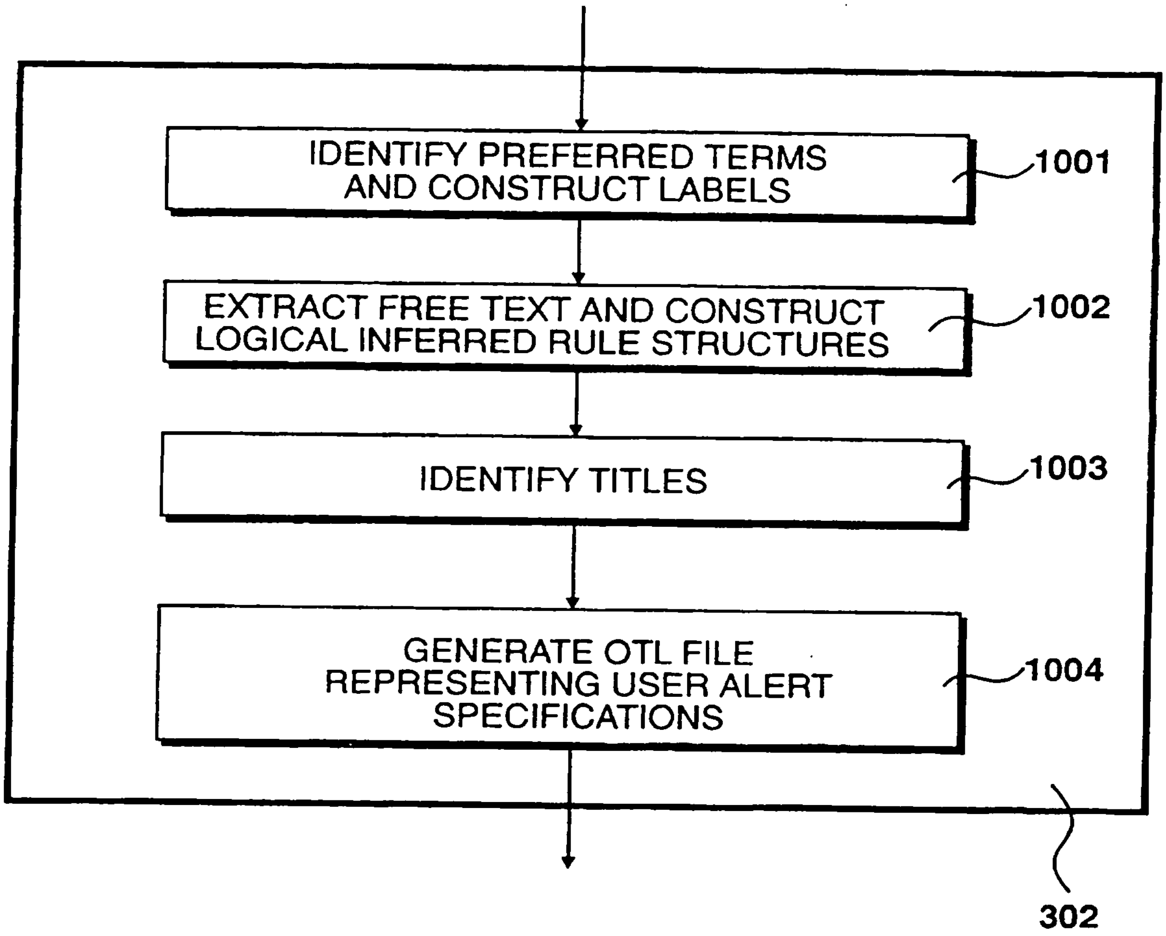
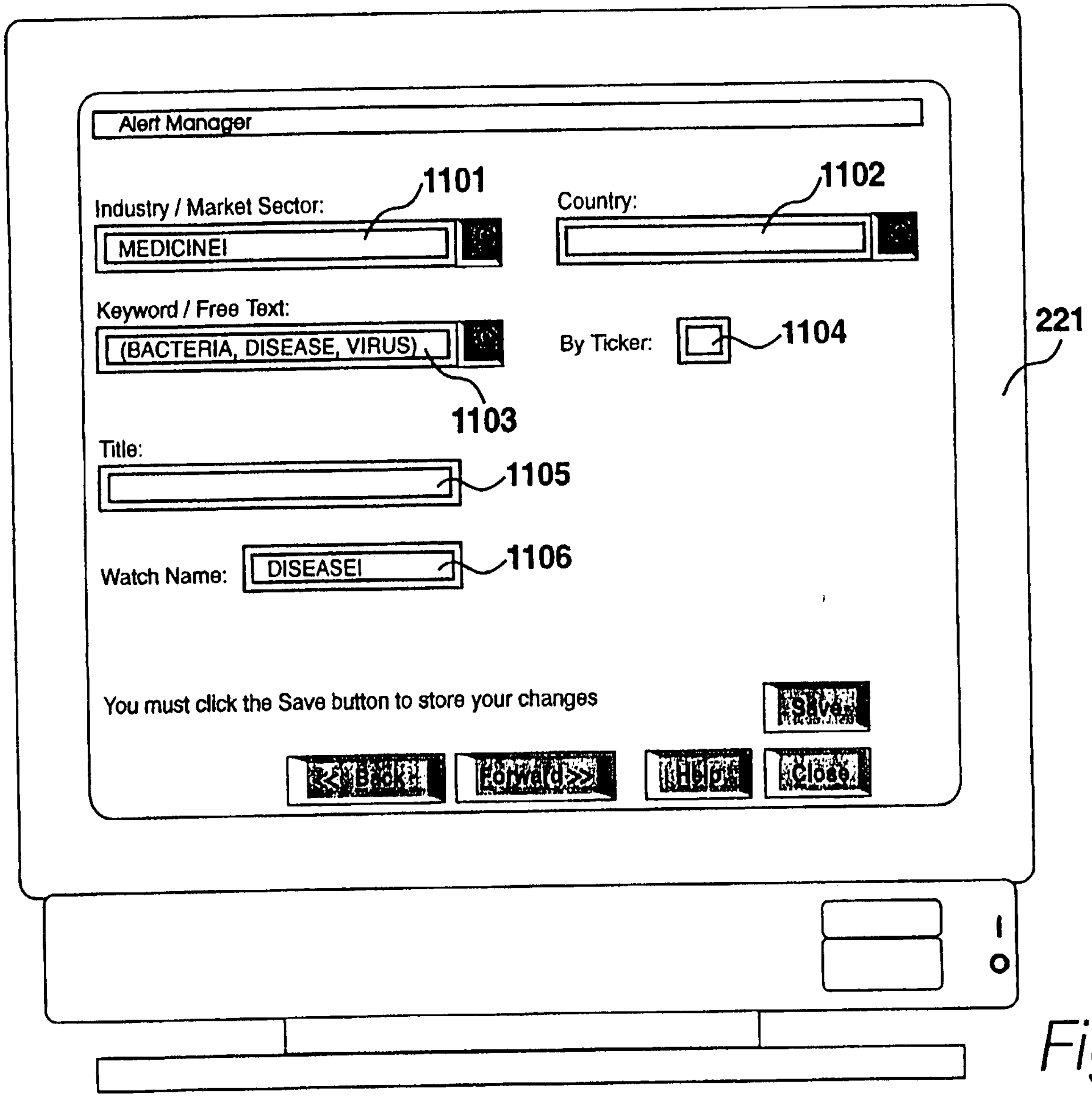

IDENTIFY PREFERRED TERMS AND CONSTRUCT LABELS
1001
EXTRACT FREE TEXT AND CONSTRUCT LOGICAL INFERRED RULE STRUCTURES
1002
IDENTIFY TITLES
1003
GENERATE OTL FILE REPRESENTING USER ALERT SPECIFICATIONS
1004
302

Figure 10



Alert Manager

Industry / Market Sector: MEDICINEI

Country:

Keyword / Free Text: (BACTERIA, DISEASE, VIRUS)

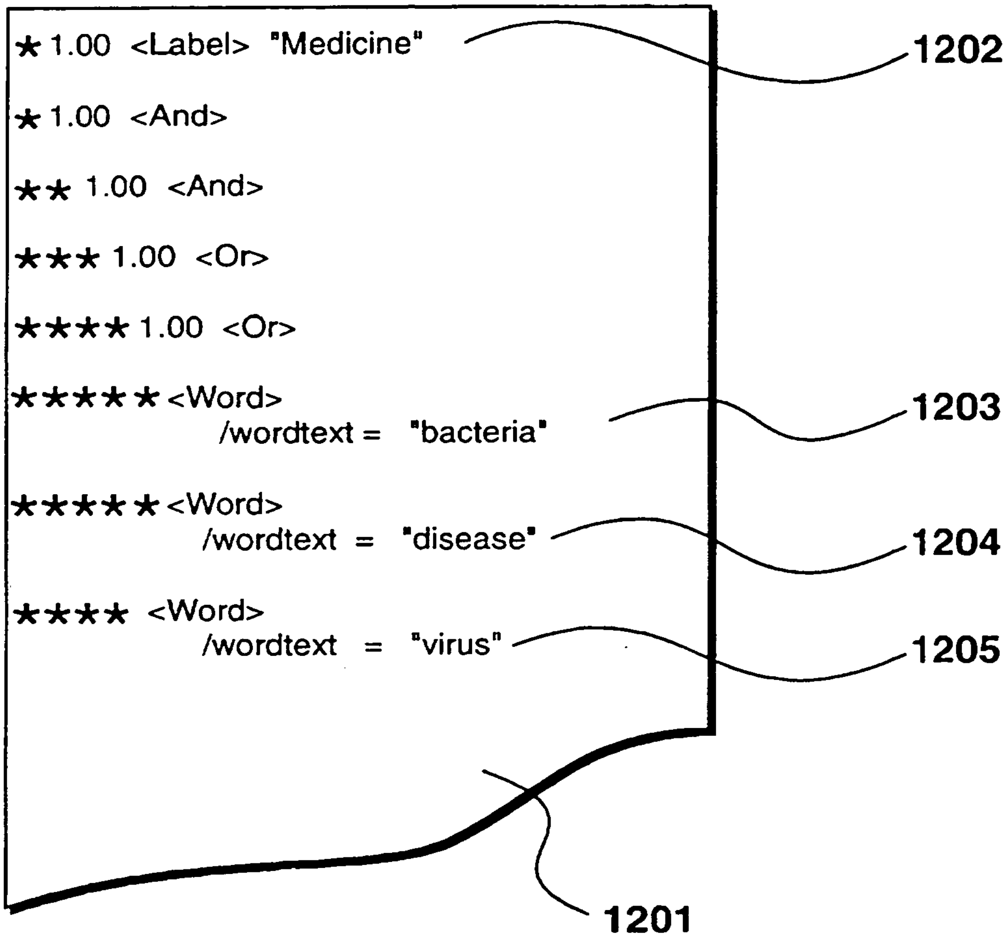
By Ticker: □

Title:

Watch Name: DISEASEI

You must click the Save button to store your changes

Figure 11



```
★ 1.00  <Label>  "Medicine"

★ 1.00  <And>

★★ 1.00  <And>

★★★ 1.00  <Or>

★★★★ 1.00  <Or>

★★★★★ <Word>
              /wordtext =   "bacteria"

★★★★★ <Word>
             /wordtext  =   "disease"

★★★★  <Word>
             /wordtext   =   "virus"
```

1202
1203
1204
1205
1201

Figure 12

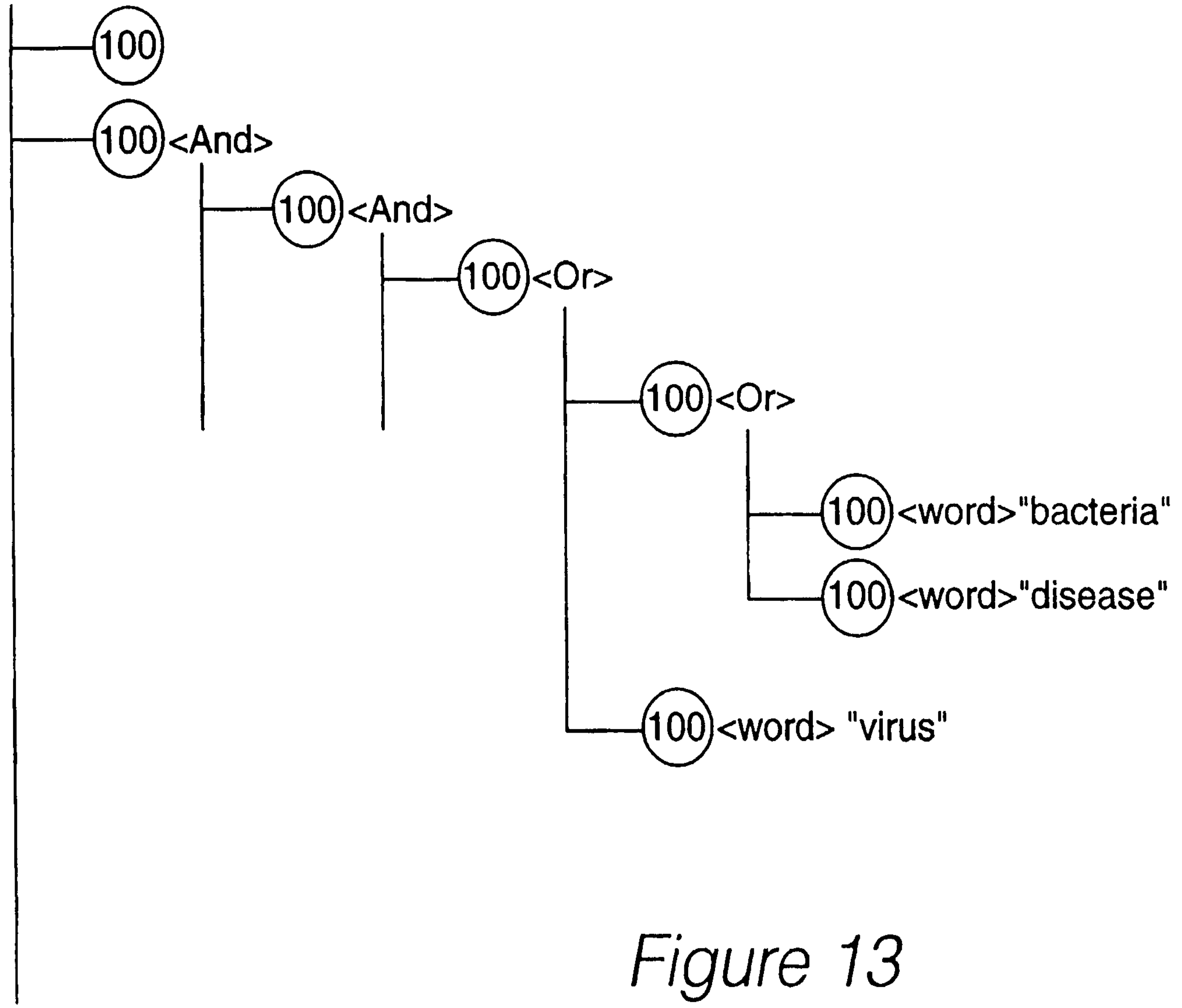
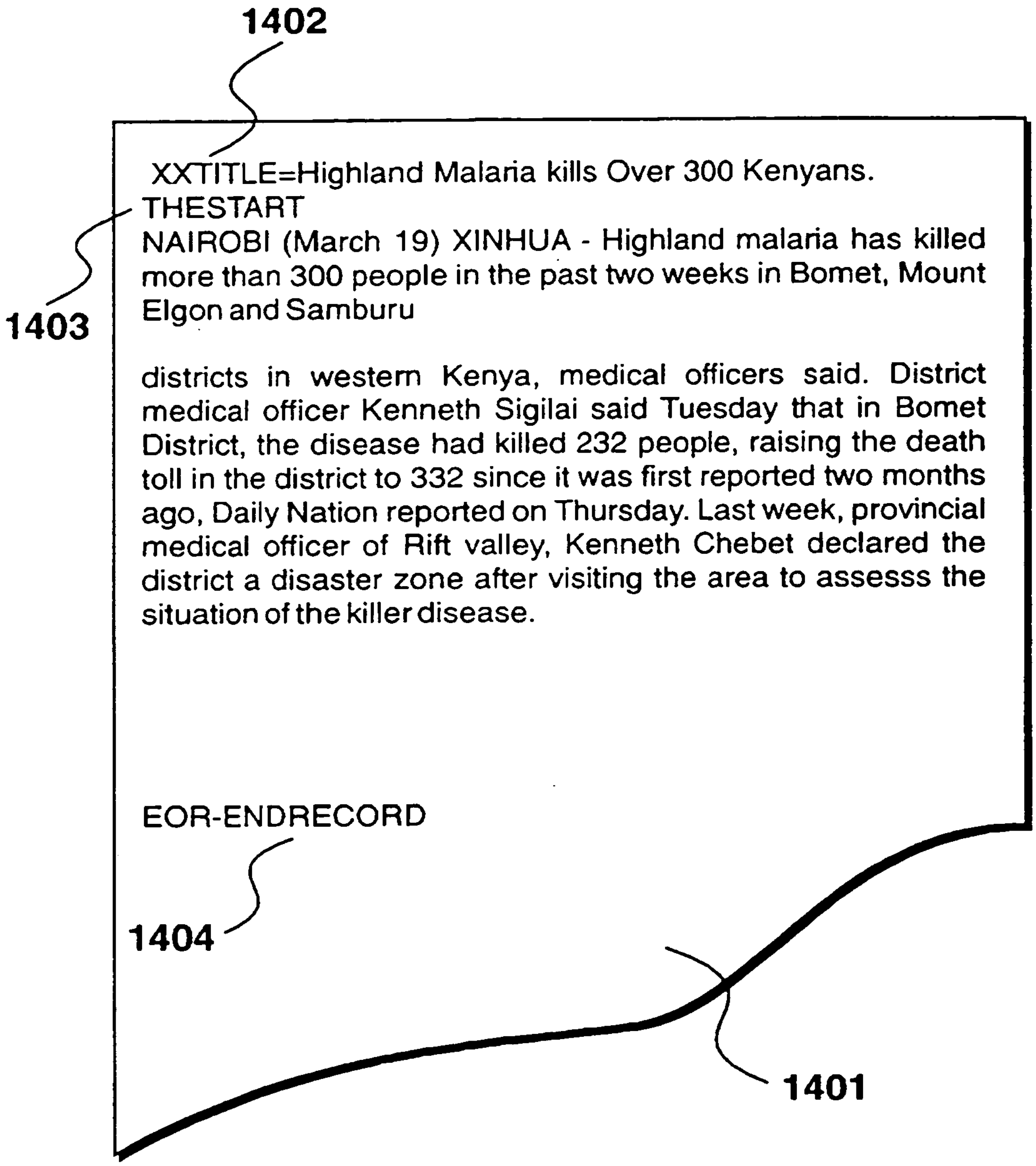


Figure 13



XXTITLE=Highland Malaria kills Over 300 Kenyans.
THESTART
NAIROBI (March 19) XINHUA - Highland malaria has killed more than 300 people in the past two weeks in Bomet, Mount Elgon and Samburu

districts in western Kenya, medical officers said. District medical officer Kenneth Sigilai said Tuesday that in Bomet District, the disease had killed 232 people, raising the death toll in the district to 332 since it was first reported two months ago, Daily Nation reported on Thursday. Last week, provincial medical officer of Rift valley, Kenneth Chebet declared the district a disaster zone after visiting the area to assesss the situation of the killer disease.

EOR-ENDRECORD

Figure 14

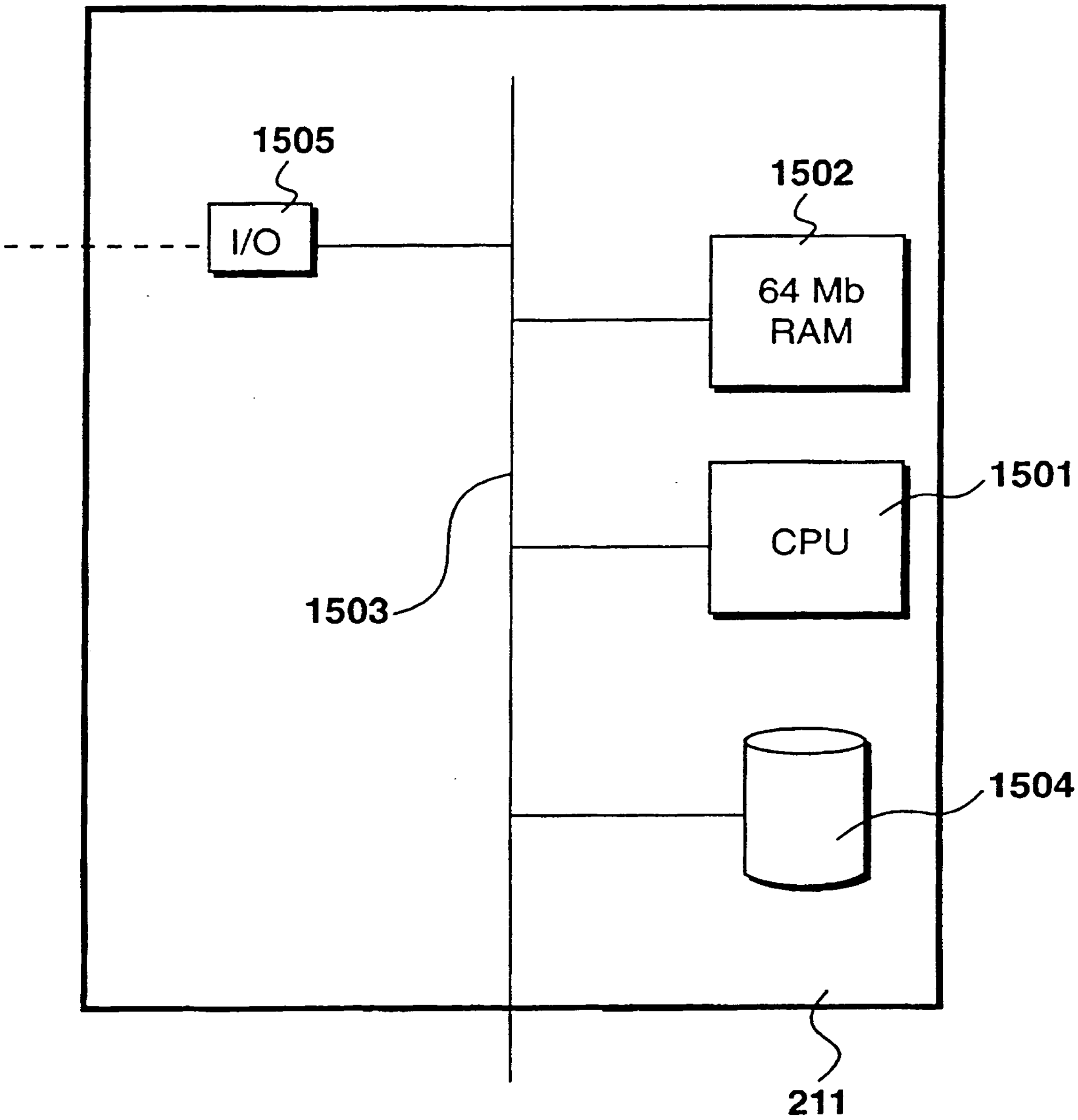


Figure 15

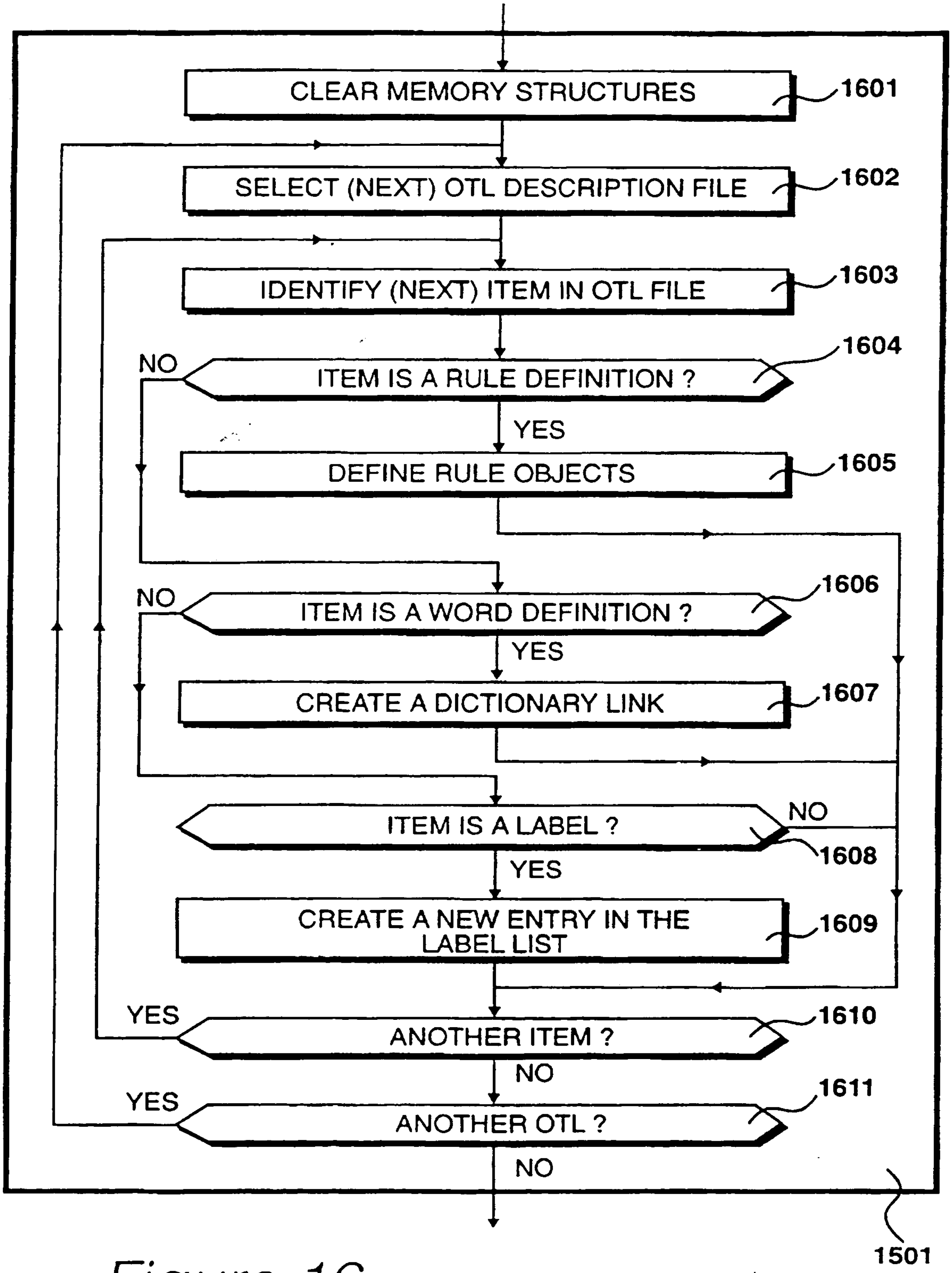

CLEAR MEMORY STRUCTURES
1601
SELECT (NEXT) OTL DESCRIPTION FILE
1602
IDENTIFY (NEXT) ITEM IN OTL FILE
1603
NO
ITEM IS A RULE DEFINITION ?
1604
YES
DEFINE RULE OBJECTS
1605
NO
ITEM IS A WORD DEFINITION ?
1606
YES
CREATE A DICTIONARY LINK
1607
ITEM IS A LABEL ?
NO
1608
YES
CREATE A NEW ENTRY IN THE LABEL LIST
1609
YES
ANOTHER ITEM ?
1610
NO
YES
ANOTHER OTL ?
1611
NO
1501

Figure 16

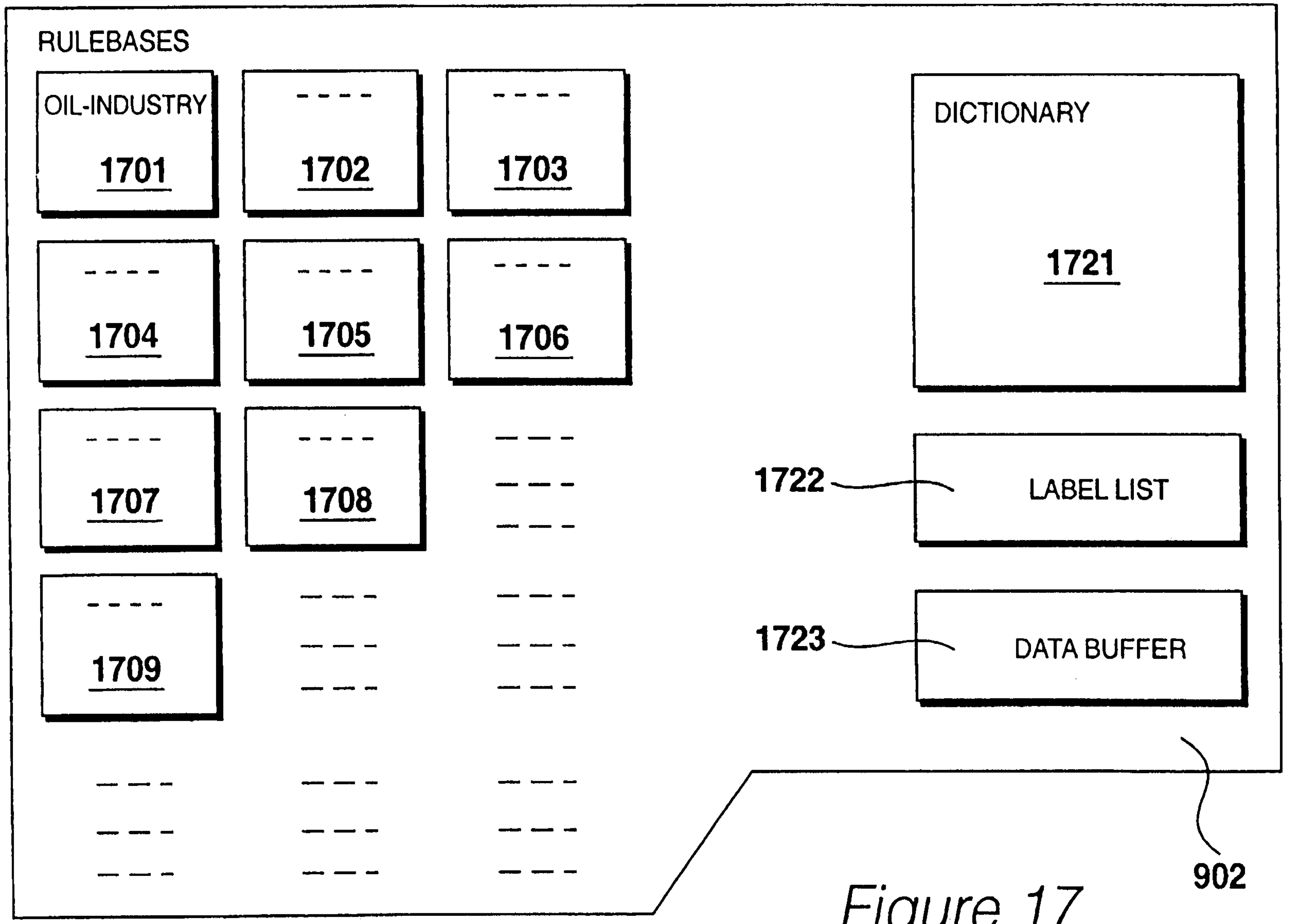

RULEBASES
OIL-INDUSTRY
1701
1702
1703
1704
1705
1706
1707
1708
1709
DICTIONARY
1721
1722
LABEL LIST
1723
DATA BUFFER
902

Figure 17

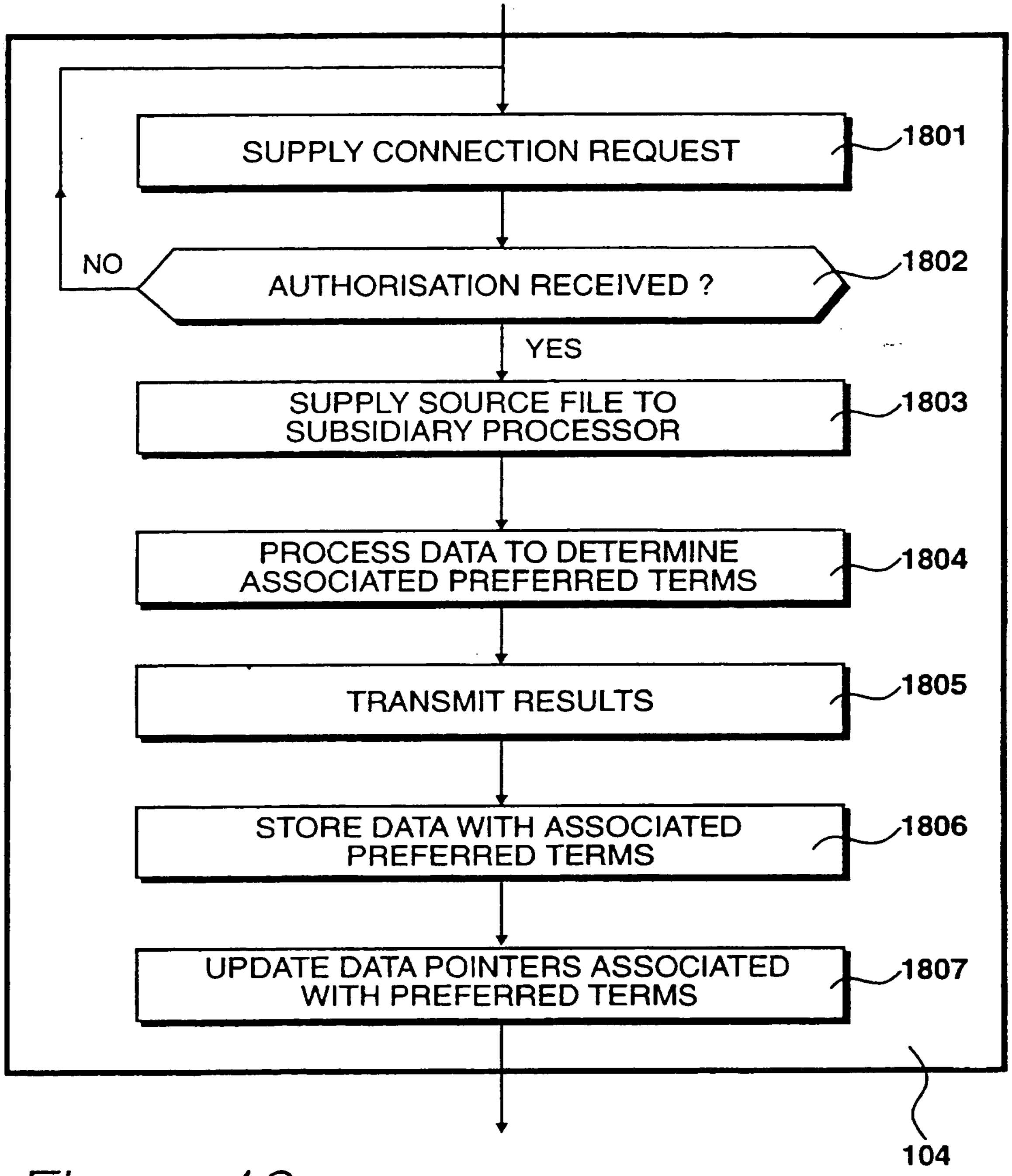

SUPPLY CONNECTION REQUEST
1801
NO
AUTHORISATION RECEIVED ?
1802
YES
SUPPLY SOURCE FILE TO SUBSIDIARY PROCESSOR
1803
PROCESS DATA TO DETERMINE ASSOCIATED PREFERRED TERMS
1804
TRANSMIT RESULTS
1805
STORE DATA WITH ASSOCIATED PREFERRED TERMS
1806
UPDATE DATA POINTERS ASSOCIATED WITH PREFERRED TERMS
1807
104

Figure 18

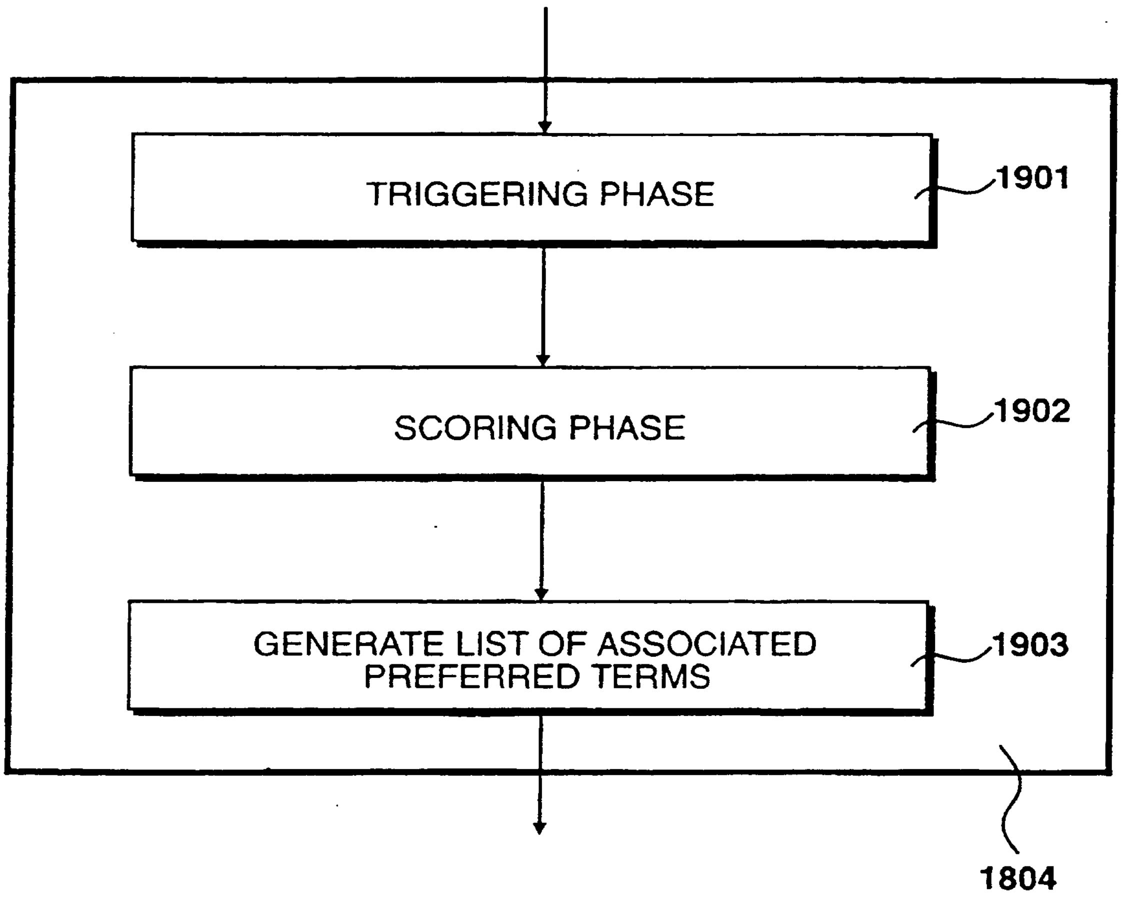

TRIGGERING PHASE
1901
SCORING PHASE
1902
GENERATE LIST OF ASSOCIATED PREFERRED TERMS
1903
1804

Figure 19

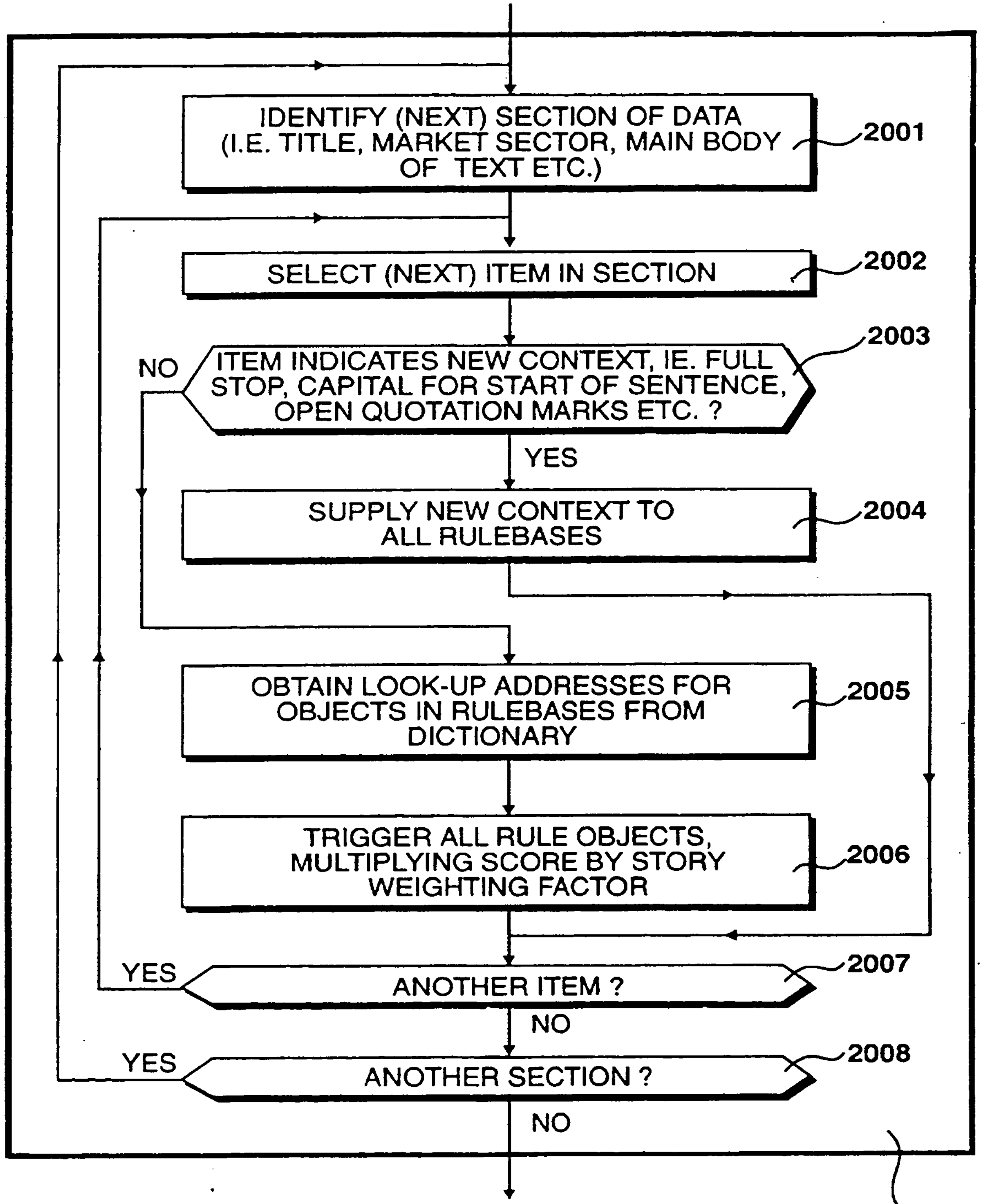

IDENTIFY (NEXT) SECTION OF DATA (I.E. TITLE, MARKET SECTOR, MAIN BODY OF TEXT ETC.)
2001
SELECT (NEXT) ITEM IN SECTION
2002
ITEM INDICATES NEW CONTEXT, IE. FULL STOP, CAPITAL FOR START OF SENTENCE, OPEN QUOTATION MARKS ETC. ?
2003
NO
YES
SUPPLY NEW CONTEXT TO ALL RULEBASES
2004
OBTAIN LOOK-UP ADDRESSES FOR OBJECTS IN RULEBASES FROM DICTIONARY
2005
TRIGGER ALL RULE OBJECTS, MULTIPLYING SCORE BY STORY WEIGHTING FACTOR
2006
YES
ANOTHER ITEM ?
2007
NO
YES
ANOTHER SECTION ?
2008
NO
1901

Figure 20

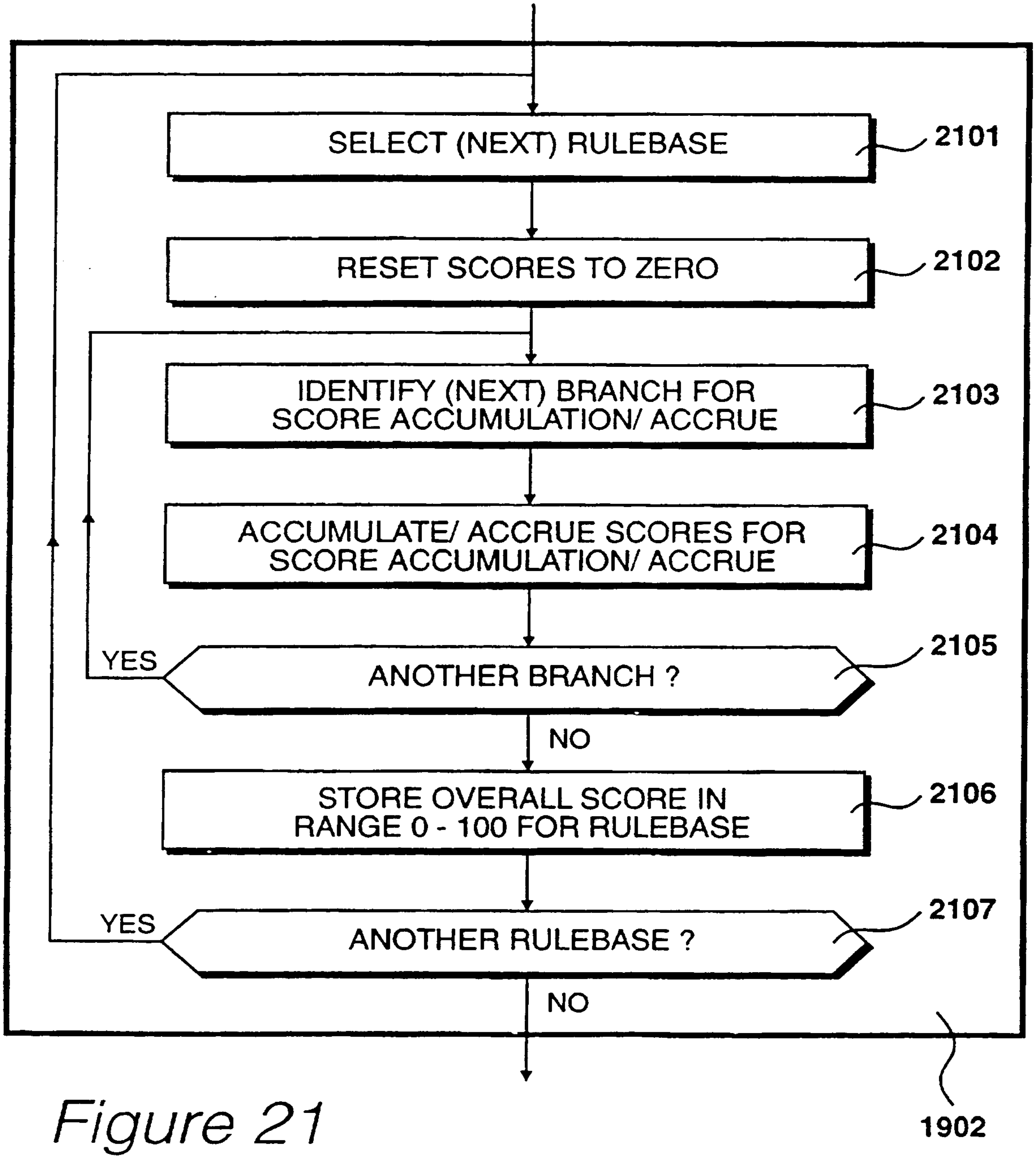

SELECT (NEXT) RULEBASE
2101
RESET SCORES TO ZERO
2102
IDENTIFY (NEXT) BRANCH FOR SCORE ACCUMULATION/ ACCRUE
2103
ACCUMULATE/ ACCRUE SCORES FOR SCORE ACCUMULATION/ ACCRUE
2104
YES
ANOTHER BRANCH ?
2105
NO
STORE OVERALL SCORE IN RANGE 0 - 100 FOR RULEBASE
2106
YES
ANOTHER RULEBASE ?
2107
NO
1902

Figure 21

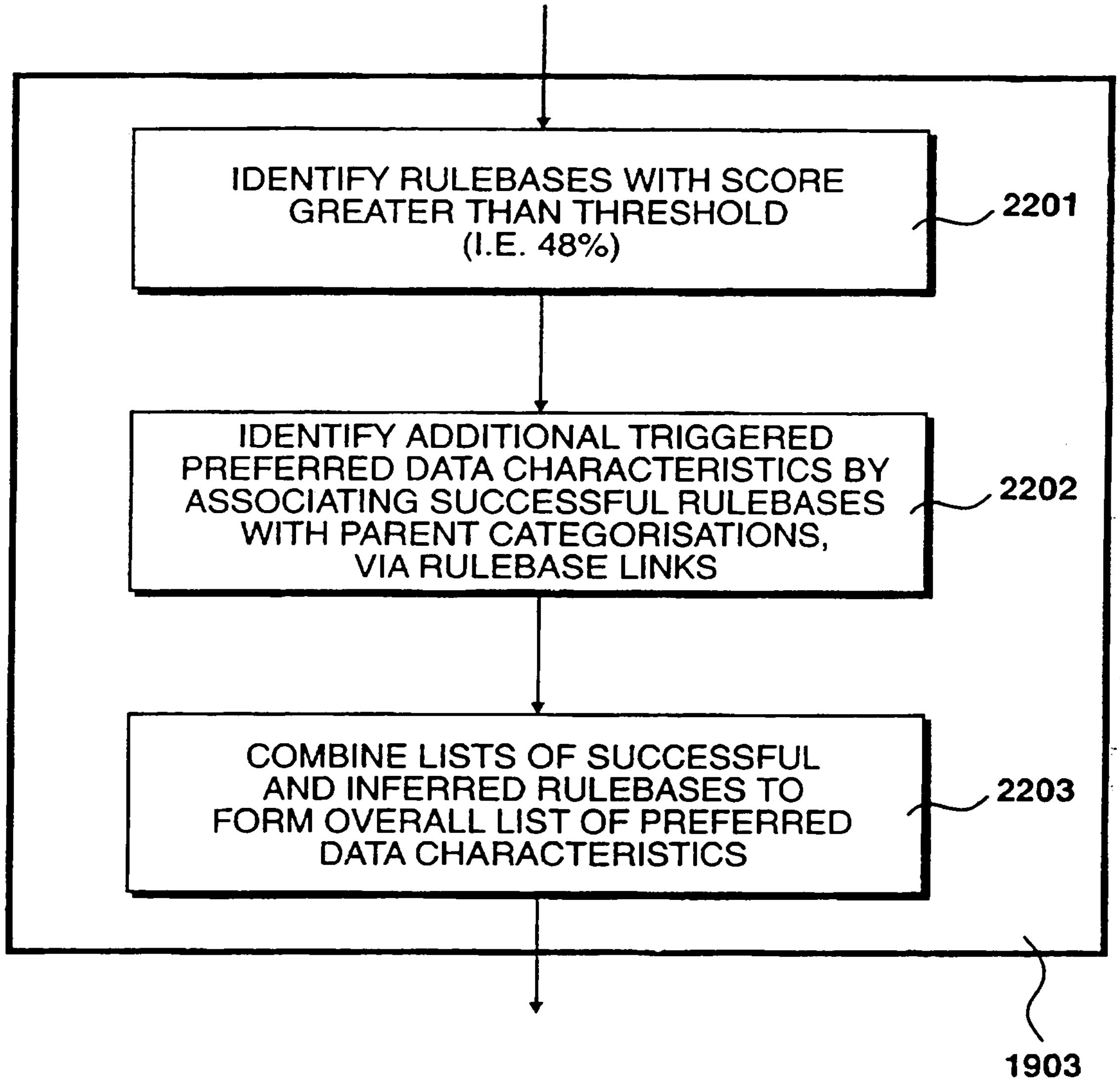
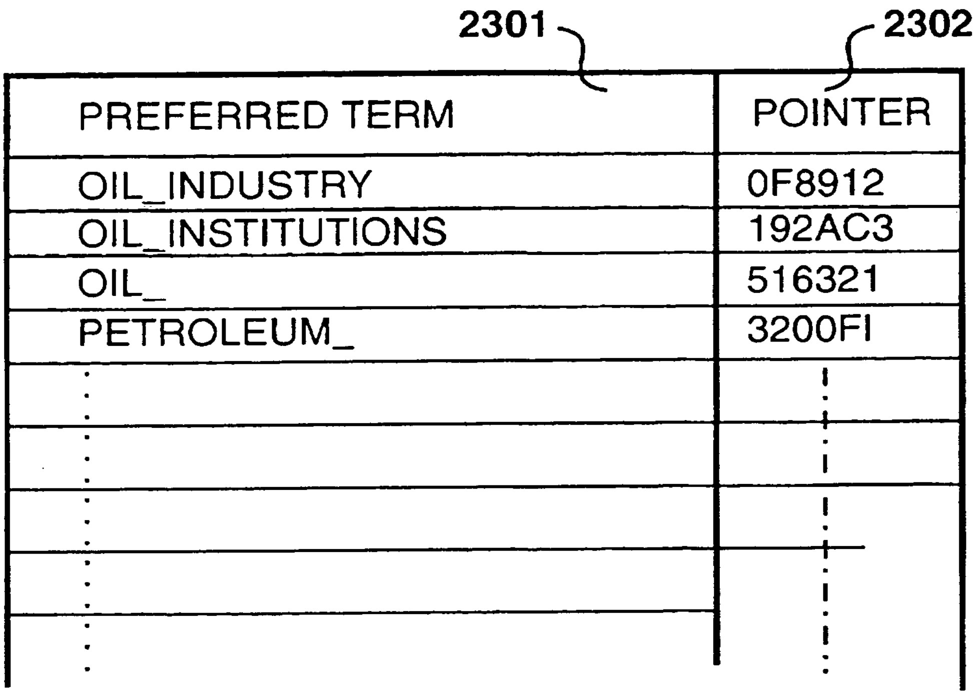
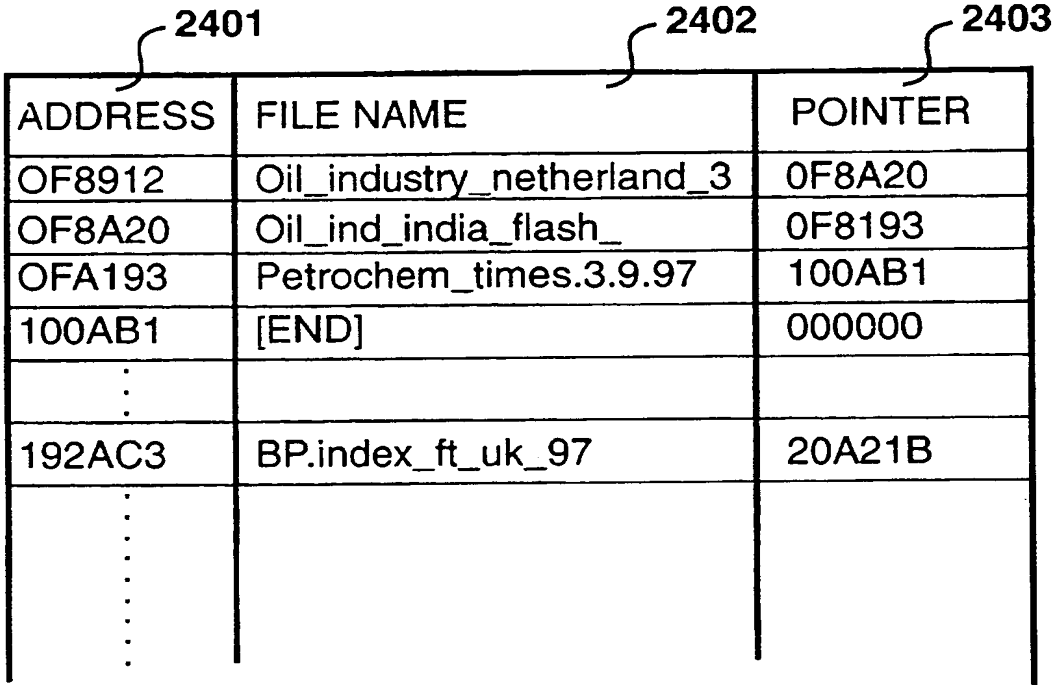

IDENTIFY RULEBASES WITH SCORE GREATER THAN THRESHOLD (I.E. 48%)
2201
IDENTIFY ADDITIONAL TRIGGERED PREFERRED DATA CHARACTERISTICS BY ASSOCIATING SUCCESSFUL RULEBASES WITH PARENT CATEGORISATIONS, VIA RULEBASE LINKS
2202
COMBINE LISTS OF SUCCESSFUL AND INFERRED RULEBASES TO FORM OVERALL LIST OF PREFERRED DATA CHARACTERISTICS
2203
1903

Figure 22



PREFERRED TERM	POINTER
OIL_INDUSTRY	0F8912
OIL_INSTITUTIONS	192AC3
OIL_	516321
PETROLEUM_	3200FI
⋮	⋮

Figure 23


2401
2402
2403

ADDRESS	FILE NAME	POINTER
OF8912	Oil_industry_netherland_3	0F8A20
OF8A20	Oil_ind_india_flash_	0F8193
OFA193	Petrochem_times.3.9.97	100AB1
100AB1	[END]	000000
⋮		
192AC3	BP.index_ft_uk_97	20A21B
⋮		

Figure 24

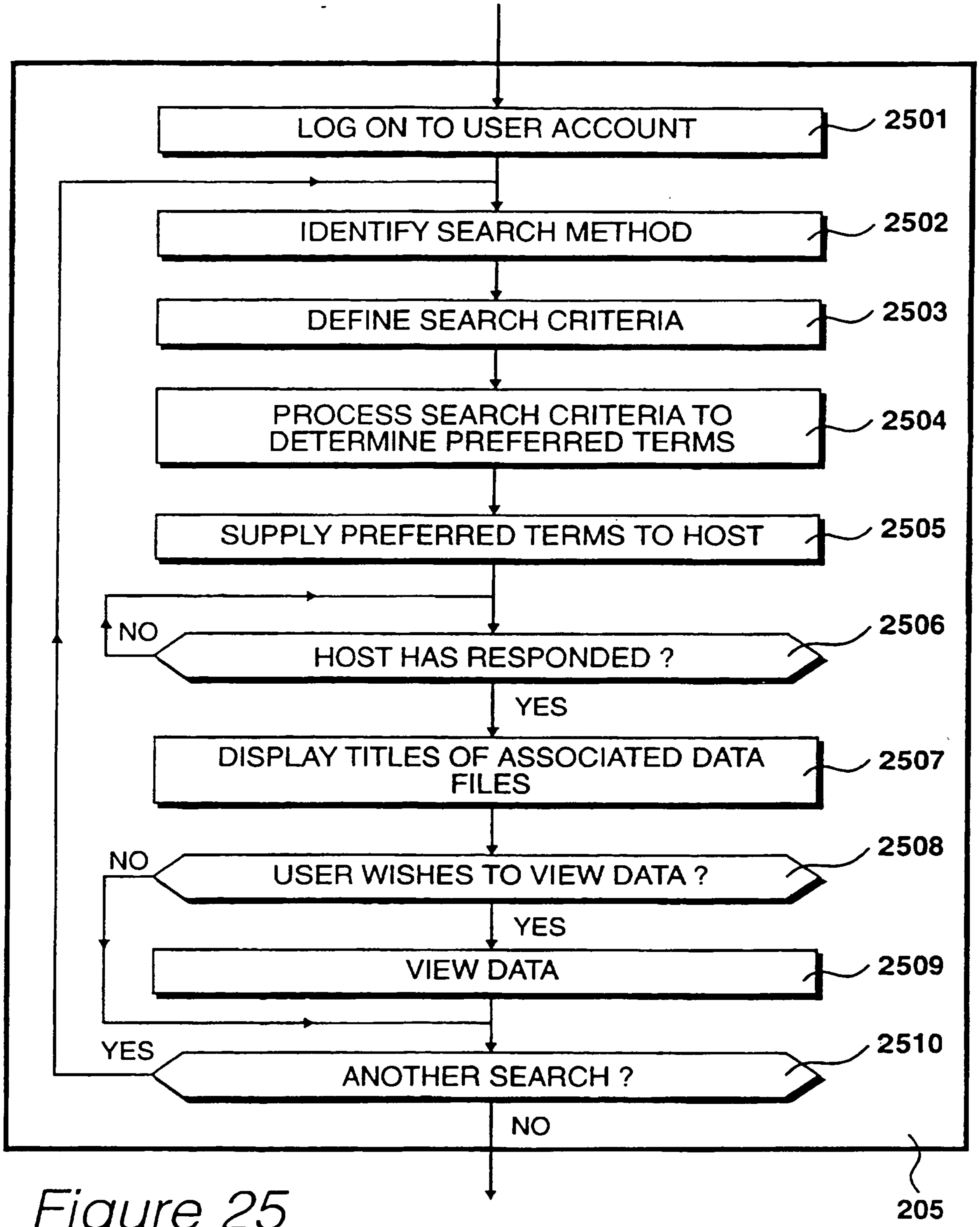
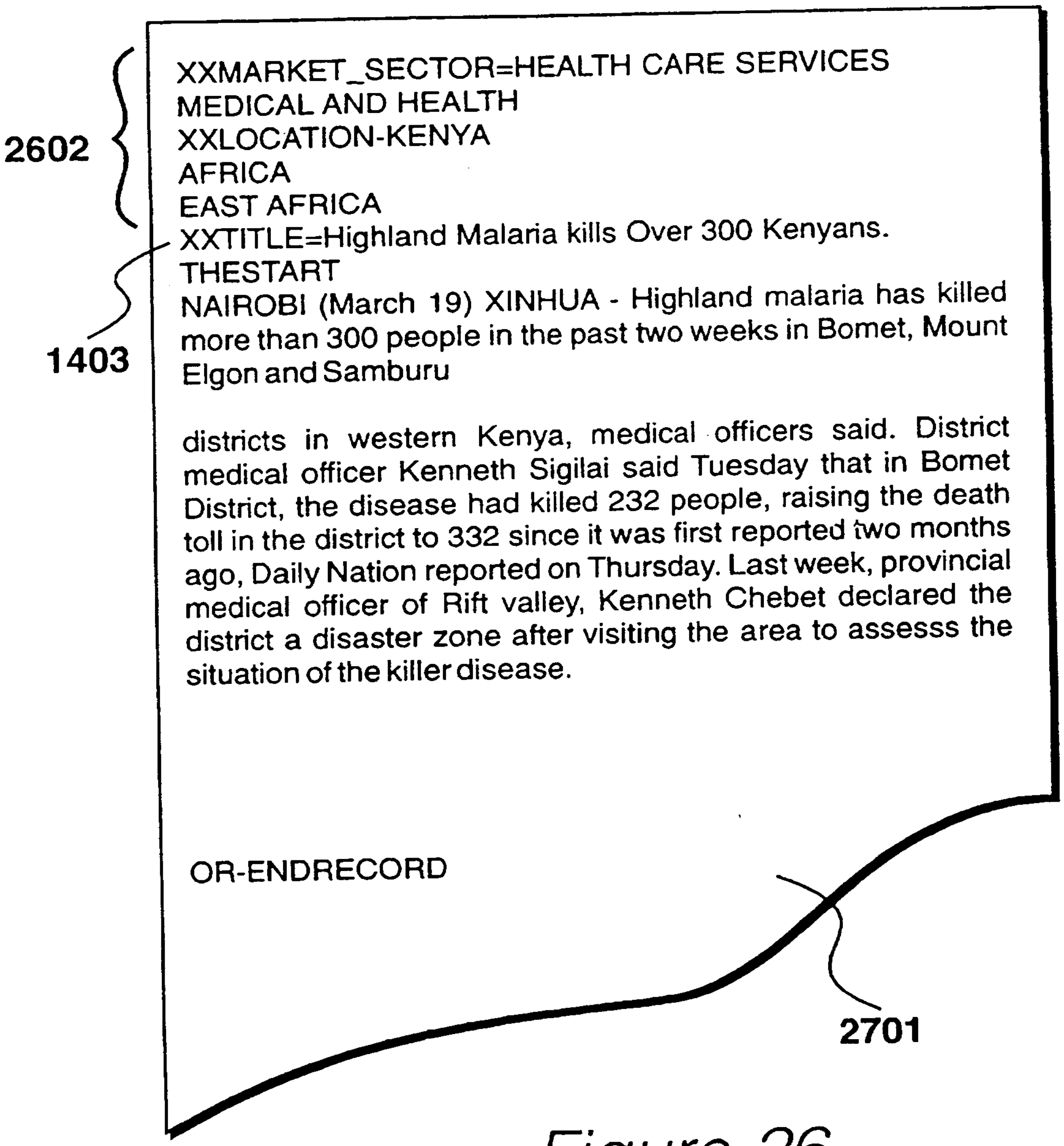

LOG ON TO USER ACCOUNT
2501
IDENTIFY SEARCH METHOD
2502
DEFINE SEARCH CRITERIA
2503
PROCESS SEARCH CRITERIA TO DETERMINE PREFERRED TERMS
2504
SUPPLY PREFERRED TERMS TO HOST
2505
NO
HOST HAS RESPONDED ?
2506
YES
DISPLAY TITLES OF ASSOCIATED DATA FILES
2507
NO
USER WISHES TO VIEW DATA ?
2508
YES
VIEW DATA
2509
YES
ANOTHER SEARCH ?
2510
NO
205

Figure 25



XXMARKET_SECTOR=HEALTH CARE SERVICES
MEDICAL AND HEALTH
XXLOCATION-KENYA
AFRICA
EAST AFRICA
XXTITLE=Highland Malaria kills Over 300 Kenyans.
THESTART
NAIROBI (March 19) XINHUA - Highland malaria has killed more than 300 people in the past two weeks in Bomet, Mount Elgon and Samburu

districts in western Kenya, medical officers said. District medical officer Kenneth Sigilai said Tuesday that in Bomet District, the disease had killed 232 people, raising the death toll in the district to 332 since it was first reported two months ago, Daily Nation reported on Thursday. Last week, provincial medical officer of Rift valley, Kenneth Chebet declared the district a disaster zone after visiting the area to assesss the situation of the killer disease.

OR-ENDRECORD

Figure 26

2336697

ALERTING USER-PROCESSING SITES AS TO THE AVAILABITLITY OF INFORMATION

The present invention relates to the alerting of user-processing sites as to the availability of data associated with user-specific characteristics.

Introduction

Search engines are known for identifying particular text files of interest from large, often distributed, databases. These known processes operate by performing free text searching in which a user specifies words which they believe are contained within the target file.

A problem with this known technique is that a simple enquiry can generate thousands of hits many of which are totally irrelevant to the user's needs. Furthermore, many relevant files may be missed because they do not actually contain the specific words chosen.

Procedures for classifying volumes of data so as to facilitate subsequent searching are known but the classification process often involves manual intervention thereby making it time consuming and prone to human error. Procedures are known for processing data files so as to determine whether the file should be associated with a particular information category. The known processes require machine readable association files (or outline files) which are used as a basis for analysing an incoming data file. The processing of a data file in combination with an outline results in a numerical score value being produced, defining an extent to which the data file is relevant to a particular category. Thereafter, a decision may be made as to whether the file should be included in the category by a threshold comparison.

In practical systems, thousands of such outline files would be required in order to provide a useful level of categorisation. The applicant's co-pending

British patent application (DGC-P11-GB) describes a procedure for automating the generation of outline files by making reference to files that have already been placed in the category and by making reference to files that are not appropriate to the category. In this way, new categories may be identified and appropriate outline files constructed.

Outline files work well if the size of incoming data files is similar to the size of the files referred to during the outline file generation process. If an incoming data file is smaller than the preferred size, score values may be adjusted as described in the applicant's co-pending British patent application (DGC-P12-GB). Alternatively, if an incoming data file is much larger than the preferred size, the file may be divided into a plurality of file sections whereafter the categorisation process is performed for each of the sections, as described in the applicant's co-pending British patent application (DGC-P13-GB).

Some categories may be considered as being particularly important to users and users may wish to receive immediate notification upon particular files being detected. Although many categories may be included within a system, these categories may be less than perfectly adequate in terms of defining the highly important issues of interest. Under these circumstances, specific outline files could be generated for a particular application. However, it is appreciated that such outline files are probably only of interest to particular users and the duration over which these outline files are required may be relatively short. Thus, it is difficult to justify the generation of outline files, in commercial terms, for specific user applications.

Summary of The Invention

According to a first aspect of the present invention, there is provided a method of alerting a user processing site as to the availability of data associated with user specified characteristics, comprising a first processing step of analysing incoming data with respect to common data characteristics to generate common category associations; a second processing step of analysing said incoming data with respect to user-specific data characteristics to generate user-specific associations; and a third processing step of generating an alerting signal to the effect that user-specific associations have been generated, wherein said user-specific data characteristics include examples of said common data characteristics, and said second processing step makes use of said common category associations.

In a preferred embodiment, incoming data files are analysed to make common category associations prior to making user-specific categorisations. Preferably, the common category associations are written to an association table.

In a preferred embodiment, incoming data files are analysed for alert conditions with respect to a plurality of users. Preferably, each user is invited to select common categories, to define free text, to define a title and/or to define a country of origin.

According to a second aspect of the present invention, there is provided a data processing system configured to analyse incoming data files and to generate alert signals to a user if a data file is detected as being relevant to characteristics defined by a said user, comprising first processing means for analysing incoming data files with respect to common data characteristics to generate common category associations; second processing means configured to analyse said incoming data files with respect to user-specific characteristics to generate user-specific associations; and a third processing means for generating an alerting signal to the effect that user-specific associations have been generated, wherein said second processing means analyses user-specific data characteristics having referenced common category associations defined by said first processing means.

In a preferred embodiment, a communication link is maintained between the processing system and a user system and relaying means are configured to relay an alert condition upon detection of said condition.

Preferably, memory means are configured to store outline association files, wherein the processing means is configured to analyse data files with reference to the outline files.

Brief Description of The Drawings

Figure 1 shows a data distribution environment in which data is received from a plurality of data sources;

Figure 2 details a data processing, storage and retrieval system shown in *Figure 1*, including a central processing system, a user specific processor and a plurality of subsidiary processors;

Figure 3 identifies procedures performed by the data procession, storage and retrieval system shown in *Figure 1*;

Figure 4 details the process for generating common characteristics for association with data files identified in *Figure 3*;

Figure 5 details the process for generating or modifying an outline file identified in *Figure 4*;

Figure 6 shows a terminal display of outline files represented graphically;

Figure 7 details an outline file from which the display shown in *Figure 6* is generated;

Figure 8 shows a diagrammatic representation of the file data shown in *Figure 7*;

Figure 9 details process **302** for the generation of user-specific characteristics;

Figure 10 details the process identified in *Figure 9* for generating an alert outline file;

Figure 11 shows a visual display at a user terminal, inviting a user to provide input information;

Figure 12 shows an example of an outline file representing user alert specifications;

Figure 13 represents a structure derived from the file shown in *Figure 12*;

Figure 14 shows an example of a source data file;

Figure 15 details a subsidiary process shown in *Figure 2*;

Figure 16 details operations performed by the subsidiary process detailed in *Figure 15*;

Figure 17 shows a plurality of rulebases produced by the process shown in *Figure 16* and stored in the memory identified in *Figure 15*;

Figure 18 details procedures performed by the data processing system **104** in response to receiving a new data file;

Figure 19 details procedures for the processing of data to determine associated preferred terms shown in *Figure 18*;

Figure 20 details a triggering phase identified in *Figure 19*;

Figure 21 details a scoring phase identified in *Figure 19*;

Figure 22 details a list generation phase identified in *Figure 19*;

Figure 23 details a table constructed by the central processing system shown in *Figure 2*;

Figure 24 details a linked list;

Figure 25 details procedures for performing a search in response to a user request;

Figure 26 shows an example of a common data associated file.

Detailed Description of The Preferred Embodiments

The invention will now be described by way of example only with reference to the above identified drawings.

A data distribution environment is illustrated in *Figure 1* in which data, received from a plurality of data sources **101**, **102**, **103** is supplied to a data processing, storage and retrieval system **104**. Data sources **101** and **102** supply data directly to processing system **104** while data source **103** supplies data via a local area network **105**, thereby allowing user terminals **106** and **107** to gain direct access to their local data source **103**.

The processing system **104** provides access to a plurality of users, such as users **111**, **112**, **113**, **114**, **115**, **116** and **117**. User **111** has direct access to the processing system **104** while users **112**, **113** and **114** gain access to the processing system **104** via the Internet **118**. Users **115**, **116** and **117** exist within a more sophisticated environment in which they have access, via a local area network **119** to their own local database system **120** in addition to a connection, via an interface **121**, to the data processing system **104**.

All incoming data from data sources **101** to **103** is categorised with a key word in seven separate fields, comprising "market sector", "location", "company name", "publisher", "publication date" and "scope". Users, such as users **112** to **117** may specify almost any term as the basis for a search and are then prompted by an equivalent word or phrase which constitutes more preferred search parameters. For example, a user may specify a search word such as "confectionery" and the system will prompt the user to consider narrower terms such as "chocolate" along with related terms such as "cakes" or "desserts", or broader terms such as "food". From a simple request, a user is given an option of focusing further or of taking a broader overview of the subject under consideration.

The scope of an article refers to the context in which the document or article was written. For example, the scope field may consider questions as to whether the article concerns "mergers and acquisitions" or "seasonal trends" et cetera. Such terms are useful in gathering related information from a wide variety of industries and markets and may prove invaluable for particular applications.

Processing system **104** is detailed in *Figure 2*. Data signals from data sources **101** to **103** are supplied to input interfaces **201** via data input lines **202.** Similarly, output data signals are supplied to users **111** to **117** via an output interface **203** and output wires **204**. Input interface **201** and output interface **203** communicate with a central processing system **205** based on DEC Alpha integrated circuitry. The central processing system **205** also communicates with other processing systems in a distributed processing architecture. Processing system **104** includes eight Intel chip based processing systems **211** to **218**, each implementing instructions under the control of conventional operating systems such as Windows NT.

An operator communicates with the processing system **104** by means of an operator terminal, having a visual display unit **221** and a manually operable keyboard **222**. Data files received from sources **101** to **103** are written to bulk storage devices **223** in the form of large magnetic disk arrays. Data files are written to disk arrays **223** after these files have been associated with preferred terms, as illustrated at step **203**. These association processes are performed by the subsidiary processors **211** to **218** and the central processing system **205** is mainly concerned with the switching and transferring of data between the interface circuits **201**, **203** and the disk arrays **223**.

The central processing system **205** communicates with the subsidiary processors **211** to **218** via an Ethernet connection **206** and processing requirements are distributed between processors **211** to **218**. Having addressed a subsidiary processor **211** to **218** the transferring of data to an addressed processor is performed. Each individual incoming data file is supplied exclusively to one of the subsidiary processors. The selected subsidiary processor is then responsible for performing the association process, to identify preferred terms relevant to that particular data file. Thereafter, the associated data file is returned to the central processing system **205**, over connection **206** and the central processing system **205** is then responsible for writing the associated data file to the disk array **223**. In this way, it is possible to scale the degree of processing capacity provided by system **104** in dependence upon the volume of data files to be processed in this way. The central processing system **205** also maintains a table of preferred terms, pointing to particular data files which have been identified as relevant to said preferred terms.

A new incoming data file is supplied to central processing system **205** from input interface **201**. The central processing system **205** supplies the new data file to one of the subsidiary processors **211** to **218** over network connection **206**. The selected subsidiary processor performs a first processing step of analysing the incoming data file with respect to common data characteristics to generate common category associations. These common category associations, which may be identified by preferred terms, effectively associate the file with particular categories thereby allowing the file to be identified with reference to these categories.

The associated common data characteristics or preferred terms are added to the file which is then returned back to the central processing system **205**. Central processing system **205** also maintains a table **228** recording details of particular associated files for each of the common categories. Thus, given a particular common category it is possible to identify all associated files and given a particular file it is possible to identify the particular common categories under which that file has been associated.

In addition to associating the files to common categories, the system shown in *Figure 2* is also capable of associating files to user-specific data characteristics. Such characteristics are defined by users, such as users **112** to **117** and the association process is performed by user-specific processor **226**. After identifying common data characteristics, as a first process, the central processing system **205** supplies data files to the user-specific processor **226** so as to allow said processor to perform a second processing step. Under the secŏnd processing step the incoming data file is analysed with respect to user-specific data characteristics to generate user-specific associations. Such associations, when identified, are brought to the attention of the central processing system **205**. The central processing system is then prompted to perform a third processing step consisting of the generation of an alerting signal to the effect that user-specific associations have been generated.

In order to obtain maximum benefit from the first processing step performed by the subsidiary processors **211** to **218** and in order to minimise the burden placed on the user-specific processor **226**, user-specific data characteristics may include examples of the common data characteristics. Furthermore, the second processing step makes use of these specified common data characteristics and relies upon processing procedures performed as part of the first processing step by the subsidiary processors **211** to **218**. Thus, an association process is performed only once, either in accordance with the common category associations, in response to operations performed by subsidiary processors **211** to **218**, or in response to user-specific operations under control of user-specific processor **226**.

Procedures performed by the data processing system **104** are summarised in *Figure 3*. Steps **301** and **302** represent set-up procedures performed prior to receiving incoming data files. Steps **303** to **307** represent the on-line procedures configured to respond as incoming data files are received. Furthermore, it should be appreciated that other procedures are performed in a multi-tasking environment, possibly in response to incoming data files, although not essential to the present invention.

At step **301** common characteristics are generated for association with data files. These common characteristics are determined by the service provider and will be established in an attempt to anticipate the demands of users.

At step **302** user-specific characteristics are generated for association with data files. These user-specific characteristics will be determined by the specific requirements of a particular user therefore, in a working environment, many user-specific characteristic sets will be created enabling the requirements of many users to be satisfied.

After generating common characteristics and user-specific characteristics, the system enters its on-line mode of operation initiated by step **303**. At step **303** a question is asked as to whether a source file has been received and when answered in the negative the system enters a short wait state at **304** before addressing the question again at step **303**. When a source file is received the question asked at step **303** is answered in the affirmative and control is directed to step **305**.

At step **305** common characteristics are associated with the incoming file and a question is then asked at step **306** as to whether any associations have been made at step **305**. If this question is answered in the affirmative, the associations identified at step **305** are written to an association table at step **307** and the file is stored by storage device **223** with the details of the associations.

At step **308** a file of user characteristics is selected and at step **309** the user characteristics selected at step **308** are associated to the received file. At step **310** the question is asked whether any associations have been made and if answered in the affirmative an alert signal to this effect is generated at step **311**. Alternatively, step **311** is bypassed to direct control to step **312**.

At step **312** a question is asked as to whether another set of user characteristics are to be considered and when answered in the affirmative control is returned to step **308**. Thus, in this way, all of the user sets are considered and alert signals are generated where appropriate. Eventually, all of the user characteristics will have been considered and control will be directed to step **313**.

At step **313** a question is asked as to whether characteristics are to be set up and when answered in the affirmative, control is returned to step **301**, effectively taking the system off-line and allowing common characteristics to be modified at step **301** or user-specific characteristics to be modified at step **302**. However, in a multi-tasking environment, it should be appreciated that it would be possible to perform the off-line and on-line functionality simultaneously. If the question asked at step **313** is answered in the negative, to the effect that on-line processing is to continue, control is returned to step **303** to await the next incoming file.

Process **301** for specifying preferred terms for association with data files is detailed in *Figure 4*. At step **401** a preferred term is selected and at step **402** an outline (OTL) file is generated or modified. At step **403** a question is asked as to whether another term is to be processed and when answered in the affirmative control is returned to step **401**, allowing the next term to be processed at step **402**. Eventually, all of the terms will have been processed resulting in appropriate generations or modifications to their related outline files. Consequently, the question asked at step **403** is answered in the negative whereafter at step **404** data structures are initialised by parsing the OTL files generated at step **402**.

Step **402** for the generation or modification of outline files is detailed in *Figure 5*. At step **501** a visual OTL editor is opened resulting in the editor's visual interface being displayed on VDU **321**. At step **502** a question is asked as to whether an existing file is to be loaded for modification and if answered in the negative a new OTL file is created at step **503**. If the question asked at step **502** is answered in the affirmative, step **503** is bypassed and at step **504** modifications or additions are made to the OTL definition. At step **505** the OTL modifications created at step **504** are tested on a sample of test data and at step **506** a question is asked as to whether another modification is to be made. When answered in the affirmative, control is returned to step **504** resulting in further modifications or additions being made to the OTL definitions. When answered in the negative at step **506**, the new or modified OTL file is saved at step **507**.

When performing modifications or additions at step **504**, a graphical representation of the OTL file data is presented to an operator via the visual display unit **321**. An example of a display of this type is illustrated in *Figure 6*, representing a graphical illustration of a specific OTL file.

The OTL file stores definitions in an hierarchical tree structure and this structure is represented in the graphical view as shown in *Figure 6*. A representation of the tree may be contracted or expanded and the possibility of expanding a particular branch is identified by a plus sign on a particular line, as shown at **601**. Similarly, when a particular branch has been fully expanded, the line is identified by a minus sign as shown at **602**. Definitions within the file consist of rules, words and labels. The labels allow relationships to be defined between various parts of the file and between individual files themselves. The words identify specific words within an input file of interest and the rules define how and what weights are to be attributed to these words. Each rule line includes, at its beginning, a weight value **603** representing the score that will be attributed when a particular rule condition is met. Rules may also have leaves and the rule defines the way in which scores generated from leaves are combined.

OTL file data represented graphically in the form shown in *Figure 6* is actually stored in a data file having a format of the type shown in *Figure 7*. The actual data file shown in *Figure 7* corresponds to the data display in *Figure 6* but in *Figure 7* all of the data, some of which has been rolled up in *Figure 6*, is present. The data contained within the file shown in *Figure 7* is manipulated interactively by an operator in response to the graphical interface displayed as illustrated in *Figure 6*. Score values **603** are also identified in the data file shown in *Figure 7*.

Displayed line **601** in *Figure 6* is generated from line **701** of the actual stored data. The syntax of the language used for recording the data, as illustrated in *Figure 7*, may vary and the example shown is specific to this particular application. However, the underlying functionality of the language may be considered with reference to the diagrammatic representation shown in *Figure 8*.

Purely to provide a specific example, this particular outline file is concerned with the topic of the oil industry and therefore the purpose of the OTL file is to identify words and phrases within an input file so as to provide an indication as to how relevant that input data is to users having an interest in the oil industry. Thus, the purpose of procedures exploiting these OTL files is to generate evidence showing that a particular data file conveys information which may be of interest to those studying the oil industry.

The outlines analyse data files in order to produce numerical evidence as to the relevance of a particular file with relation to a particular topic. The OTL definitions and structures are determined empirically and would be modified and upgraded over a period of time. The system does more than merely register the existence of a particular word item by placing the word items within an interacting structure; the nature of which is illustrated in *Figure 8*. The particular entry, given label "oil-industry-mkt" relates to marketing aspects of the oil industry and as such can contribute to an overall score as to the pertinence of incoming data to this particular topic. The first line **801** shows that this particular contribution may provide a total score of forty percent. This total of forty percent is then subdivided such that at line **802** the presence of the phase "buying oil from" has a score of fifty percent. Thus, the total contribution made the presence of this phrase consists of fifty percent of forty percent, that is a total of twenty percent being made to the total contribution. Similarly, as shown at line **803** and below, particular words may be identified which result in contributions of sixty percent of thirty percent of forty percent. Thus, a complete OTL file is structured in this way with particular words and phrases making contributions to an overall score value. These words and phrases may also be specified in the rules as making single contributions or being allowed to accrue.

Process **302** for the generation of user-specific characteristics is detailed in *Figure 9*. At step **901** a user is invited to select common categories of the type specified at step **201**. At step **902** a user is invited to define user-specific data characteristics which may be in the form of key words or free text. At step **903** a user is invited to define a specific file title and at step **904** a user is invited to specify a particular country of origin.

At step **905** the user is invited to define a particular alert format, specifying the way in which the user is alerted when a new data file has been received which satisfies the user's data characteristics. At step **906** the user definitions are processed to generate an alert outline (OTL) file.

The user's alert criteria include components, defined at step **902**, which require extensive searching of new material as it is received. Searching of this type places a significant burden upon the information supplying resource. In addition, the characteristics also include reference to the preferred terms which will have been associated automatically upon receiving each data file by means of the subsidiary processors **311** to **318**. In accordance with the present preferred embodiment, the user-specific characteristics include examples of common data characteristics, specified at step **901** and reference to these characteristics are included in the user definitions generated at step **906**. However, when implementing these definitions, use is made of the previously processed common category associations, thereby significantly reducing the processing overhead placed on the user-specific processor **226**.

Process **906** for generating an alert OTL file is detailed in *Figure 10*. At step **1001** preferred terms are identified and labels are constructed. At step **1002** free text entries are extracted and logical inferred rule structures are constructed. At step **1003** titles are identified and at step **1004** an OTL file is generated representing the user's alert specifications. Thus, these specifications may include references to common data characteristics in combination with references to user-specific data characteristics. User **117** communicates with the data processing station **104** via a terminal including a Visual Display Unit (VDU) **221** and a manually operable keyboard **222**.

VDU **221** is shown in *Figure 11*, having received an initial screen of data from the data processing station **104**, inviting the user to provide input information in accordance with the procedures identified in *Figure* 9. Common categories may be entered within displayed boxes **1101**, **1102**, **1103**, **1104**, **1105** and **1106**. Box **1101** allows an industry or market sector to be selected, while box **1102** allows a particular country of interest to be selected. Items entered at boxes **1101** and **1102** represent common categories and allow information to be supplied back to the central system **104** in response to prompt **901**.

Keywords or free text are entered, as user-specific data characteristics, in box **1103** a specific title, as prompted by step **903**, may be entered in box **1105** and an alert format is defined by box **1104**. In this example, a user may receive an alert as an e-mail message or, alternatively, a user maintains a continuous connection with the system and the user is continually updated with alerts in a manner similar to known ticker tapes. In addition, a user may identify a particular watch name, for the particular characteristics being defined, allowing a plurality of searching procedures to run simultaneously, at box **1106**.

After supplying information into the boxes of the display shown in *Figure 11*, the information is supplied back to the central system **104**, thereby allowing processes **1001** to **1003** to be performed as detailed in *Figure 10*. This is then followed by the generation of the OTL file at step **1004**; a process performed by central processing system **205**.

Operation of step **1004** results in the production of an OTL file and an example of such a file is given in *Figure 12*. OTL file **1201** has been generated in response to the input data illustrated in *Figure 11*. Common data characteristics, such as the characteristic "medicine" entered at box

1101 is recorded in the OTL file as a label, as illustrated at line **1202**.

Asterisks beneath this show levels of nesting and effectively represent the importance of a particular phrase or relationship within the structure of the definition. Thus, below the top level label, five levels of nesting are included before a specific word is defined at line **1203**. At line **1203** word texts are derived from the free field **1103** which, in this example, result in three lines being included; the first being the word "bacteria" at line **1203**, the second being the word "disease" at line **1204** and the third being the word "virus" at line **1205**.

Figure 12 represents an example of an OTL file for a specific user's application. It is used to associate particular text files as being relevant and consistent with the search criteria supplied by the user. The file includes reference to common data characteristics in combination with reference to user-specific data characteristics. Each common data characteristic has its own OTL file, of the type illustrated in *Figure 7*. Thus, when implemented, OTL file **1201** directly performs an association process with respect to the three word-text words shown at lines **1203**, **1204** and **1205**.

The OTL file also includes examples of the common data characteristics and as such it effectively calls an existing OTL file generated for those specific common characteristics. Thus, in this way, it is not necessary to generate new OTL files for the common characteristics and it is not necessary to perform an additional search based on these characteristics, given that association processes will have already taken place. Thus, OTL file **1201** provides a sophisticated level of functionality without being required to generate significant amounts of OTL structuring because it refers to the existing OTL files for the common category associations.

The outline structure defined by file **1201** is illustrated in *Figure 13*. This structure is substantially similar to the structure of common category associations, as illustrated in *Figure 8*.

Source data files are received at step **303** and an example of a source data file is shown in *Figure 14*. All incoming data files are converted into a standard format of the type shown for file **1401**. The file includes a title identifier at **1402** taking the form "XXTITLE". This is followed by the actual title of the file followed by a delimiter "THESTART" at **1403**. The end of the body text is identified at **1404** by the string "EOR=ENDRECORD".

Upon receiving file **1401**, a central processing system **204** supplies this file to a subsidiary processor, such as subsidiary processor **211**. The subsidiary processor analyses the file with respect to common data characteristics to generate common category associations. These are added to the file itself and also recorded in table **328** before the data file is then written to storage **223**.

Subsidiary processor **211** is detailed in *Figure 15*. The processor includes an Intel Pentium processing unit **1501** connected to sixty-four megabytes of randomly accessible memory **1502** via a PCI bus **1503**. In addition, a local disk drive **1504** and an interface circuit are connected to bus **1503**. Interface circuit **1505** communicates with the TCP/IP network. Random access memory **1502** stores instructions executable by the processing unit **1501**, in addition to storing input data files received from the data sources **101** to **103** and intermediate data. Operations performed on processing unit **1501**, in response to instructions read from memory **1502** are identified in *Figure 16*.

At step **1601** temporary memory structures are cleared and at step **1602** an OTL description file is selected. At step **1603** an item in the OTL file is identified and at step **1604** a question is asked as to whether the item selected at step **1603** is a rule definition. If this question is answered in the affirmative, a rule object is defined at step **1605**. Alternatively, if the question asked at step **1604** is answered in the negative, to the effect that the item is not a rule definition, a question is asked at step **1606** as to whether the item is a word definition. If this question is answered in the affirmative, a dictionary link is created at step **1604**.

At step **1608** a question is asked as to whether the item is a label and when answered in the affirmative a new entry is created in a label list, whereafter at step **1610** a question is asked as to whether another item is present. After executing step **1605** or after executing step **1607**, control is directed to step **1610**.

When the question asked at step **1610** is answered in the affirmative, to the effect that another item is present, control is returned to step **1603** and the next item is identified in the OTL file. Eventually, all of the items will have been identified resulting in the question asked at step **1610** being answered in the negative. Thereafter, at step **1611** a question is asked as to whether another OTL file is present and when answered in the affirmative control is returned to step **1602** allowing the next OTL description file to be selected. Thus, this process continues until all of the OTL files have been considered resulting in the question asked at step **1611** being answered in the negative.

For each OTL file considered, by being selected at step **1602**, a rule base is generated and a plurality of such rule bases is illustrated in *Figure 17*. Thus, a first OTL file processed in accordance with the procedures shown in *Figure 10* results in the generation of a first rule base **1701**. Similarly, further iterations of the procedures shown in *Figure 7* result in the generation of rule bases **1702** to **1709**. Typically, for a specific installation, in the order of three thousand rule bases would be generated by execution of the procedures illustrated in *Figure 10*. Rule bases **1701** to **1709** are stored in memory **1502**, which also provides storage space for a dictionary **1721**, a label list **1722** and a data buffer **1723**. The dictionary stores a list of words which have importance in any of the stored rule bases. Associated with each word in the dictionary, there is at least one pointer and possibly many pointers, to specific entries in specific rule bases **1701** to **1709**. Thus, the words identified at **803** in *Figure 8* would all be included in dictionary **1721**. Entries within the dictionary **1721** are implemented upon execution of step **1607** in *Figure 16*.

Similarly, execution of step **1609**, creating a new entry in the label list, allows a label to relate to rules that are elsewhere in the tree structure.

Processes performed by the data processing system **104** for associating preferred terms with the source files are detailed in *Figure 18.* At step **1801** central processor **205** obtains access to one of the subsidiary processors **211** to **218**. The central processor then expects to receive authorisation so that communication may be effected with one of the subsidiary processors. after a connection has been established, the source file is supplied to the selected subsidiary processor at step **1803** and at step **1804** the data is processed to determine associated preferred terms.

After performing the processing at step **1804**, the results are transmitted back to the central processing system at step **1805** and at step **1806** data with associated preferred terms is stored and data pointers associated with the preferred data terms are updated at step **1807**.

Step **1804** for the processing of data to determine associated preferred terms is detailed in *Figure 19.* The overall processing is broken down into three major phases, consisting of a triggering phase at **1901**, followed by a scoring phase at **1902** followed finally by a list generation phase at step **1903**.

Triggering phase **1901** is detailed in *Figure 20.* At step **2001** a section of the data, such as its title, market sector or main body of text, is identified and at step **2002** an item of the identified section is selected. At step **2003** a question is asked as to whether the item indicates a new context, which may be considered as a grammatical marker in the form of a full stop, capital, start of a sentence or quotation marks et cetera. When answered in the affirmative new context information is supplied to all rule bases **1701** to **1709** at step **2004** and control is then directed to step **2007**.

If the question asked at step **2003** is answered in the negative, step **2004** is bypassed and a look-up address is obtained for rule objects in rule bases from the dictionary at step **2005**. Thereafter, at step **2006** all addressed objects are triggered and a multiplication of scores is effected by a score weighting factor. Thereafter, at step **2007** a question is asked as to whether another item is present and when answered in the affirmative control is returned to step **2002**.

Eventually all of the items for a selected section will have been considered resulting in the question asked at step **2007** being answered in the negative. Thereafter, at step **2008** a question is asked as to whether another section is to be considered and when answered in the affirmative control is returned to step **2001**.

At step **2001** the next section is identified and steps **2002** to **2008** are repeated. Eventually, all of the sections will have been considered and the question asked at step **2008** will be answered in the negative.

Scoring phase **1902** is detailed in *Figure 21*. At step **2101** a rule base is selected and at step **2102** a score variable is re-set to zero. At step **2103** a branch is identified for score accumulation/accrue and at step **2104** scores are accumulated or accrued from triggered rules attached to the branch. At step **2105** a question is asked as to whether another branch is to be considered and when answered in the affirmative control is returned to step **2103**. A next branch is selected at step **2103** with procedure **2104** being repeated. Eventually all of the branches will have been considered resulting in the question asked at step **2105** being answered in the negative.

At step **2106** an overall score in the range of zero to one hundred is stored for the rule base and at step **2107** a question is asked as to whether another rule base is present. When answered in the affirmative control is returned to step **2101** and steps **2101** to **2107** are repeated. Eventually, all of the rule bases will have been considered and the question asked at step **2107** will be answered in the negative.

The operations illustrated in *Figure 21* may be considered with reference to the illustration of the structure in *Figure 8*. Thus if any of the defined words at **803** are identified within the file a provisional score of one hundred will be allocated. However, the process as shown in *Figure 21*, must then ascend up the branches so that any scores lower down will be modified in response to scores higher up the structure.

Phase **1903** for the generation of a list of associated preferred terms is detailed in *Figure 22*. At step **2201** a rule base is identified having a score greater than a predetermined threshold. Thus, for a particular application a threshold may be set at forty-eight percent. At step **2202** additional triggered preferred data characteristics are identified by associating successful rule bases with parent categorisations by rule base links.

At step **2203** lists of successful and inferred rule bases are combined to form overall lists of preferred data characteristics. Step **2203** results in data being generated by a subsidiary processor, such as processor **211**, which is then supplied back to the central processing system **205.**

Central processing system **205** is responsible for constructing a table of the type shown in *Figure 23* in which an entry is present for each preferred term. The specific preferred terms are stored in column **2301** and, for each of these terms, column **2302** defines a specific pointer to a position in memory associated with the central processing system **205**. Specific data files are identified by file names and the number of files associated with each preferred term is variable, depending on the nature and the amount of input data being considered. Thus, in order for this data to be accessible quickly while optimising use of the storage capacity within the central processing system **205**, an indication of the file names is stored in the form of a linked list as illustrated in *Figure 24*.

The preferred term "OIL_INDUSTRY" has been associated to a pointer 0F8912, as shown *in Figure 23*. Address 0F8912 is the first in column **2401** of the linked list. Column **2402** identifies a particular file name and column **2403** identifies the next pointer in the list. Thus, entry 0F8912 points to a particular file with the file name "OIL_INDUSTRY_NETHERLAND_3" with a further pointer to memory location 0F8A20. At memory location

0F8A20 a new file name is provided, illustrated at column **2402** and again a new pointer is present at column **2403**. Eventually, all relevant files will have been considered and the end of the list is identified by address 000000 at the pointer location in column **2403**.

Procedures for performing a search in response to a user request are detailed in *Figure 25*. At step **2501** a user logs onto the system and at step **2502** a search method is identified. At step **2503** search criteria are defined and at step **2504** search criteria are processed to determine preferred terms. At step **2505** a list of preferred terms are supplied to the central processing system **205**.

At step **2506** a question is asked as to whether the host has responded and when answered in the affirmative titles of associated data files are displayed at step **2507**. At step **2508** a question is asked as to whether the user wishes to view identified data and when answered in the affirmative the data is viewed; after being downloaded over the communication channel, at step **2509**. At step **2510** a question is asked as to whether another search is to be performed and when answered in the affirmative control is returned to step **2502**.

Common data associated files are supplied to storage device **223** and an example of such a file is shown in *Figure 26*. File **2601** is a processed version of file **1401** and includes all the information present in file **1401**. In addition, reference to common categories have been added to the top of the file, as shown at **2602** before title line **1403**.

Claims

1. A method of alerting a user-processing site to a condition as to the availability of information associated with user-specified characteristics, comprising

a first processing step of analysing incoming data files with respect to common data characteristics to generate common category associations;

a second processing step of analysing said incoming data files with respect to user-specific data characteristics to generate user-specific associations; and

a third processing step of generating an alerting signal to the effect that user-specific associations have been generated, wherein

said user-specific data characteristics include examples of said common data characteristics, and said second processing step makes use of said common category associations.

2. A method according to claim 1, wherein incoming data files are analysed to make common category associations prior to making user-specific data categorisations.

3. A method according to claim 2, wherein said common category associations are written to an association table.

4. A method according to claim 1, wherein incoming data files are analysed for alert conditions with respect to a plurality of users.

5. A method according to claim 1, wherein each user is invited to select common categories.

6. A method according to claim 1, wherein each user is invited to define free text.

7. A method according to claim 1, wherein each user is invited to define a title.

8. A method according to claim 1, wherein each user is invited to define a country of origin.

9. A method according to claim 1, wherein an alert condition is relayed to a user by a maintained communications link.

10. A method according to claim 1, wherein data files are analysed with reference to outline association files so as to generate score values.

11. A data processing system configured to analyse incoming data files and to generate alert signals to a user if a data file is detected as being relevant to characteristics defined by said user, comprising

first processing means for analysing incoming data files with respect to common data characteristics to generate common category associations;

second processing means configured to analyse said incoming data files with respect to user-specific characteristics to generate user specific associations; and

a third processing means for generating an alerting signal to the effect that user-specific associations have been generated, wherein

said second processing means analyses user-specific data characteristics having reference to common category associations defined by said first processing means.

12. Apparatus according to claim 11, wherein said first processing means analyses an incoming data file with respect to common data characteristics before said second processing means analyses said incoming data file with respect to user-specific characteristics.

13. Apparatus according to claim 12, wherein said first processing means is configured to write said common category associations to an association table configured in memory.

14. Apparatus according to claim 11, wherein said second processing means is configured to analyse incoming data files with respect to alert conditions defined for a plurality of users.

15. Apparatus according to claim 11, including user selection means configured to invite each user to select common categories.

16. Apparatus according to claim 11, including user selection means configured to invite each user to define free text.

17. Apparatus according to claim 11, including user selection means configured to invite each user to define a title.

18. Apparatus according to claim 11, including user selection means inviting each user to define a country of origin.

19. Apparatus according to claim 11, including a maintained communications link between the processing system and a user's system, including relaying means configured to relay an alert condition upon detection of said condition.

20. Apparatus according to claim 11, including memory means configured to store outline association files, wherein said processing means are configured to analyse data files with reference to said outline files.

The Patent Office

26

Application No: GB 9808801.6
Claims searched: 1-20

Examiner: K. Sylvan
Date of search: 9 October 1998

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.P): G4A (AUDB)

Int Cl (Ed.6): G06F (17/30)

Other: Online: WPI,LISA

Documents considered to be relevant:

Category	Identity of document and relevant passage		Relevant to claims
A	US5537586	Individual, Inc. See column 1 lines 23-27 and column 5 lines 8-67.	-
A	US5428778	Office Express Pty. See column 1 line 51 to column 2 line 17, column 4 lines 22-34 and 53-58, and column7 lines 17-35 .	-

X Document indicating lack of novelty or inventive step
Y Document indicating lack of inventive step if combined with one or more other documents of same category.
& Member of the same patent family

A Document indicating technological background and/or state of the art.
P Document published on or after the declared priority date but before the filing date of this invention.
E Patent document published on or after, but with priority date earlier than, the filing date of this application.